Kankee Briefs LD Brief – March/April 2025

Resolved: The development of Artificial General Intelligence is immoral.

Lincoln Douglas 2025 March April Brief

Volunteers: Brett Boelkens (Editor), George Shubitidze, Cyris Lim, Shreeya Ram, Aalekh Arapalli, Daniel Romero, Aiden Tringali, Kallen Zhou, and Azzy (Alexa Xiang)

A colorful triangle with a black background

Description automatically generated

## Letter From The Editor

Worthwhile topic research is a hefty task for the experienced debaters and coaches – this is why many debate organizations have consolidated topic research in the form of briefs. However, debate briefs cost hundreds of dollars over the course of a year of topics, which can add up quickly, especially for smaller programs just starting off. Some of the best debaters in the country often spend thousands of dollars on debate briefs and private coaches who have competed and coached debate for decades that will strategize and research for them. This is a stark comparison to a small school, which sometimes will not even have a full-time debate coach. Debate is expensive enough without trying to outcompete the best and brightest minds money can buy.

Money should never be the reason why someone’s debate career is not successful. Kankee Briefs hopes to eliminate those financial inequities as much as possible by removing the research gap between those who can afford private coaching, and those who can’t. For the past six years, we’ve released debate briefs for every NSDA topic, at no charge to debaters.

Every two months during the school year, the most recent Kankee Brief can be sent directly to your email inbox if you sign up for our email list by filling out the Google Form linked below with your name and email.

Email List: <https://docs.google.com/forms/d/e/1FAIpQLScgb24XvQe1SAQb4med2NkCEqD6BIsO3K7ERbuoCUGdjSKVBg/viewform>

Kankee Briefs will always remain free for everyone, but we have a small favor to ask. If you found any value in what we do, please support us via Patreon. Your support helps free up our time for higher quality evidence and topic analyses where we otherwise would put hours at the office. Your support is entirely optional, but is still exceptionally appreciated by our volunteers. You can also be recognized by name in every new Kankee Brief release in the list of our supporters.

Please email me at [karkingkankee@gmail.com](mailto:karkingkankee@gmail.com) if you have any questions, comments, or concerns.

## Supporters

#### Hector Vela

#### Sanjeev Kumar

#### Jordan hart

#### Julie Larnard-Newbury

#### Devyn (Toby) Blaylock

#### Scott Brown

## Topic Analysis

### 1.1 Introduction

#### Introduction

Resolved: The development of Artificial General Intelligence is immoral.

Since the public release of ChatGPT in 2022, public discussion of artificial intelligence has been perhaps at an all-time high. I recall first accessing Gemini (then called Bard) with my father, in which we made Gemini answer impeccably stupid questions or write fanfiction about him being a cowboy. Since then, new AI tech has been developing at extraordinarily fast rates, with newfangled capabilities for AI to generate new songs, research papers, art, movies, deepfakes, etc., with eerie similarity to that otherwise created by humans.

If you were unfamiliar with Kankee Briefs, there is a non-zero possibility of you being skeptical of whether this topic analysis itself was written by AI. Someone too could compile every topic analysis released by Kankee Briefs and possibly create an AI emulation of my writing (please don’t do this).

All this goes to say that powerful AI will impact virtually every industry imaginable, potentially automating many intellectually heavy industries and possibly causing unemployment for many. Everyone has potential reasons to both take advantage of and be wary of others’ usage of AI, with massive implications for the job market, educational outcomes, military power, culture, and democracy.

However, conversation and worries have been almost exclusively in terms of large language models (LLMs), such as ChatGPT, Gemini, or Deepseek (or other similarly narrow AI). These are all AIs with particularly specialized applications. For instance, machine learning is a key tool in the creation of AI for many tool-assisted speedruns of video games (see the link [here](https://www.youtube.com/watch?v=qv6UVOQ0F44) for an explanation), discovering new exploits/glitches and leads to a super-human performance. Similarly, AI in games like chess is so vastly superior to humanity that it is impossible for us to win without a handicap.

These all are seemingly fantastical, almost techno-futurist-esque examples of a particular artificial intelligence’s advantage over humans. However, these are all particular AIs designed for a single or small subset or purposes – IBM’s Watson cannot play Super Mario World much like how ChatGPT is not a self-driving car. These are narrowly designed tools for a specific purpose with virtually no ability to operate outside its trained environment.

Artificial general intelligence (AGI) is **categorically** different from narrow AI. By almost every credible definition conceivable by experts qualified in the subject, AGI is a hypothetical future technology with potential to revolutionize economies, science, and government. Think not of chatbots, rather a potential Skynet or Ultron (note that a pro-human AGI is also possible).

Some technologists believe in the technological singularity occurring with AGI improving its own capabilities and intelligence, causing a potential “intelligence explosion” and the creation of superintelligence(s). A post-work or even a post-human future is a possibility if a singularity occurs, with AGI being the new dominant intelligence. As a matter of comparison, ChatGPT struggles with math and tells people to eat glass as medical advice.

#### AGI is distinct from narrow AI like generative AI – AGI needs to be adaptable, cross applicable, and generate new solutions with the intelligence at or above a human level

Baum and Villasenor 23 [Jeremy Baum, undergraduate student and researcher and the UCLA Institute for Technology, Law, and Policy, and John Villasenor, Nonresident Senior Fellow for Governance Studies at the Center for Technology Innovation at Brookings and professor of electrical engineering, law, public policy, and management at UCLA, 07-18-2023, "How close are we to AI that surpasses human intelligence?", Brookings, https://www.brookings.edu/articles/how-close-are-we-to-ai-that-surpasses-human-intelligence/]/Kankee

For decades, superintelligent artificial intelligence (AI) has been a staple of science fiction, embodied in books and movies about androids, robot uprisings, and a world taken over by computers. As far-fetched as those plots often were, they played off a very real mix of fascination, curiosity, and trepidation regarding the potential to build intelligent machines. Today, public interest in AI is at an all-time high. With the headlines in recent months about generative AI systems like ChatGPT, there is also a different phrase that has started to enter the broader dialog: artificial general intelligence, or AGI. But what exactly is AGI, and how close are today’s technologies to achieving it? Despite the similarity in the phrases generative AI and artificial general intelligence, they have very different meanings. As a post from IBM explains, “Generative AI refers to deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on.” However, the ability of an AI system to generate content does not necessarily mean that its intelligence is general. To better understand artificial general intelligence, it helps to first understand how it differs from today’s AI, which is highly specialized. For example, an AI chess program is extraordinarily good at playing chess, but if you ask it to write an essay on the causes of World War I, it won’t be of any use. Its intelligence is limited to one specific domain. Other examples of specialized AI include the systems that provide content recommendations on the social media platform TikTok, navigation decisions in driverless cars, and purchase recommendations from Amazon. AGI: A range of definitions By contrast, AGI refers to a much broader form of machine intelligence. There is no single, formally recognized definition of AGI—rather, there is a range of definitions that include the following: Source Definition of Artificial General Intelligence (AGI) OpenAI’s charter “…highly autonomous systems that outperform humans at most economically valuable work” Hal Hodson, in The Economist “[a] hypothetical computer program that can perform intellectual tasks as well as, or better than, a human.” Gary Marcus “…any intelligence (there might be many) that is flexible and general, with resourcefulness and reliability comparable to (or beyond) human intelligence.” Sébastien Bubeck et al. “…systems that demonstrate broad capabilities of intelligence, including reasoning, planning, and the ability to learn from experience, and with these capabilities at or above human-level.” While the OpenAI definition ties AGI to the ability to “outperform humans at most economically valuable work,” today’s systems are nowhere near that capable. Consider Indeed’s list of the most common jobs in the U.S. As of March 2023, the first 10 jobs on that list were: cashier, food preparation worker, stocking associate, laborer, janitor, construction worker, bookkeeper, server, medical assistant, and bartender. These jobs require not only intellectual capacity but, crucially, most of them require a far higher degree of manual dexterity than today’s most advanced AI robotics systems can achieve. None of the other AGI definitions in the table specifically mention economic value. Another contrast evident in the table is that while the OpenAI AGI definition requires outperforming humans, the other definitions only require AGI to perform at levels comparable to humans. Common to all of the definitions, either explicitly or implicitly, is the concept that an AGI system can perform tasks across many domains, adapt to the changes in its environment, and solve new problems—not only the ones in its training data. GPT-4: Sparks of AGI?

#### Current AI is not AGI – generative AI and large language models aren’t AGI

Sublime 24 [Jeremie Sublime, Associate Professor at the ISEP with a PhD in Applied Computer Science from AgroParisTech and a Master of Engineering Degree in Computer and Information Technologies from INHA University in South Korea, 01-10-2024, “The AI Race: Why Current Neural Network-based Architectures are a Poor Basis for Artificial General Intelligence,” Journal of Artificial Intelligence Research, https://dl-acm-org.ezproxy.library.unlv.edu/doi/pdf/10.1613/jair.1.15315]/Kankee

However, we have also discussed the limits of these systems, and how what appears to look like intelligence is sometimes an illusion made possible by the vast amounts of data ingested by these systems:

 First, none of these systems is nowhere near being generic. While they are somewhat more versatile than their predecessors, they remain limited to a narrow range of tasks: the fact that we have 3 categories of AIs discussed in this paper is proof enough of the lack of general purpose.

 While assistant AIs appear to have reasoning and common sense abilities, they fail on a very regular basis to solve very simple problems, which shows that this illusion of reasoning and common sense is mainly due to brute force learning. The same could be said of gaming AIs that play moves that are difficult to explain for a humans, but sometimes fail to see that the game is over and they have lost in situations where it would be impossible to miss for a human.

 The idea of creativity is very subjective and difficult to evaluate, even for humans. We could argue that most of what appears to be creative with these systems is merely the statistical result of very long adversarial learning sessions with a bit of randomness here and there. And what to say about the hallucinations that LLMs systems have when they make up things because they have to answer something ? Is it creativity, or a proof of stupidity ?

 Finally, we have seen that the adaptability and genericity of these systems is severely hindered by the neural network architectures that power them: They are limited in terms of how they can interface with the world, and their ability to evolve is constrained by an architecture that needs to be retrained if it is modified.

 We have also shown that this same architecture currently prevents the explosion of intelligence, the singularity that some scientists from fields outside of artificial intel- ligence appear to fear.

Some scientists even go as far as saying that any illusion of emergent abilities in the latest versions of these systems (that is LLMs) can easily be debunked with the right statistical tools and enough time (Schaeffer, Miranda, & Koyejo, 2023). In short, we can say that despite undeniable and very impressive progresses, current AIs systems are very far from achieving AGI. We could even argue that they are not really in the right direction because of the limitations imposed by neural network based architectures in terms of volumes of data required to train them, interfacing constraints, lack of symbolic learning abilities and how these models are trained. 4.2 Challenges, Dangers and Opportunities

The distinction between (narrow) AI and AGI is immensely important – the letter “G” is critical for topic discussions regarding existential risks from superintelligence or the ethical implications of creating machine beings with potential consciousness and moral value.

One cannot simply search for cards regarding AI, as that would be inclusive of affs premised on claiming the immoral status of the likes of chess robots, deepfakes, and self-driving cars. Under that interpretation of AGI, a Roomba aff could exist (potentially citing the video linked [here](https://youtu.be/GWS9RqfC_Rw)), or an AI music aff critiquing the existence of arguably excellent AI generated music (linked [here](https://www.youtube.com/watch?v=wPlOYPGMRws)). Apologies in advance for anyone having suffered from the experience of viewing either of those links. This also means the intellectual property rights violations from AI (see links [here](https://www.tomshardware.com/tech-industry/artificial-intelligence/meta-staff-torrented-nearly-82tb-of-pirated-books-for-ai-training-court-records-reveal-copyright-violations) or [here](https://mashable.com/article/openai-microsoft-deepseek-trained-stolen-data)) don’t have a direct link to the topic.

Smart teams ought to look out for key phrases that indicate a card would be about narrow AI, as opposed to general AI. Some terms to look out for are the names of large language models (such as ChatGPT, Gemini, or Deepseek), or the generic terms like large language model, generative AI, and/or narrow AI.

However, don’t treat the lack of the word “general” in a card as a catch-all excuse to dismiss potential evidence. In topic literature, some cards referencing AGI refer to it as AI. AI is a general term and is inclusive of both narrow and general intelligences. For example, existential risk literature is near-exclusively about AGI, not narrow AI, but most literature treats AGI risks as synonymous to AI risks. AGI is a more technical definition and not everyone is familiar with the idiosyncrasies of AI literature.

A fringe potential argument is the potential implications of AGI in the context of simulations. Given *huge* influence of the author Nick Bostrom on topic literature, there is a high probability of the existence of arguments that reference his other work regarding whether we live in a simulation (i.e. Matrix).

Given his assumptions, Bostrom believes in a large likelihood of us being artificial general intelligences ourselves, as we would be artificial simulations of biological beings in a simulation, as opposed to existing in absolute reality as biological beings. There are also implications of the creation of an AGI within a simulation, such as its potential capabilities of breaking, crashing, or escaping the simulation (see the link [here](https://philarchive.org/rec/YAMHTS-2) for a more in-depth explanation). If all of this sounds immensely esoteric and strange, that is because it is, but that is par for the course on the yes/no Skynet topic. Ultimately, if we do live in a simulation and qualify as artificial general intelligences, voting aff would potentially imply that our existence is immoral and that we ought not have been created (or continue to exist).

In response, it’s worthwhile to answer with a definition (see below) regarding AGI being nonbiological (hence the term artificial), but this begs the question of whether we are artificial or biological beings (given that we would be biological in the context of the simulation, but ultimately we would be artificial simulations of biological beings).

#### AGI is only nonbiological, algorithmic system – expert consensus definitions

**Pascal and Lazar 24** [Alexander Pascal, Senior Fellow at the Ash Center for Democratic Governance and Innovation and a Professor of Practice at the Fletcher School of Law and Diplomacy at Tufts University, and Seth Lazar, a Professor of Philosophy at the Australian National University, Distinguished Research Fellow of the University of Oxford Institute for Ethics in AI, fellow of the Carnegie Endowment for International Peace, and a Senior AI Advisor to the Knight First Amendment Institute at Columbia University, 3-28-2024, “AGI and Democracy,” Harvard ASH Center for Democratic Governance and Innovation, https://ash.harvard.edu/resources/agi-and-democracy/]/Kankee

Defining AGI But what even is AGI? Defining it sometimes feels like pinning Jell-O to a wall. But as progress accelerates, something like a consensus is emerging. **Synthesising a vast literature**, we can say that AGI would be a nonbiological computational system that can perform **any** cognitive function currently performed by humans at the level of the median human or better (acknowledging the crude quantification this implies). Google DeepMind’s recent paper mentions “linguistic intelligence, mathematical and logical reasoning, spatial reasoning, interpersonal and intra-personal social intelligences, the ability to learn new skills and creativity.” Other AI researchers would also add instrumental rationality, causal reasoning, tool use, and at least some ability to distinguish truth from falsehood. OpenAI calls it, simply, “AI systems that are generally smarter than humans.”

#### Definitions of Key Terms

Prior discussions of AGI covered merely its existence and ontology - its worthwhile to discuss its potential (lack) of a matchup with human values. AGI researchers call the issue of creating an AGI that values humanity’s continued existence and promotes our flourishing as the alignment problem. Is it on our side (aligned) or its own (misaligned)? If a misaligned AGI pursues its own goals and cares not for humanity, humans will be a means to an end (or an obstacle to that end) as opposed to an end within itself.

Examples of a misaligned AGI could be Skynet (Terminator/Fallout 2), Hal 9000 (Space Odyssey), and SHODAN (System Shock) - they have independent wills aiming to accomplish their goals, but their goals are apathetic or oppositional to humanity

None of this means that alignment produces positive values. An AGI that has been militarized in order to have an allegiance to a specific country could be endowed with values that produce bad outcomes for humanity but remain aligned with the intent of its creator. For instance, in Clone Wars, Separatist droids are at war with the Republic and commit war crimes, but they are aligned (albeit against the Republic’s interests). In real life, this could mean an authoritarian country using AGI for repression or to topple democracies, or a rogue non-state actor making potentially genocidal AGI for religious extremist reasons.

Similarly, an aligned AGI with its own will could come about creative, but unethical solutions to problems. GOTO from KOTOR II is a pseudo-aligned superintelligence with a patriotic dedication to protect and preserve the Republic. He ends up creating a galactic crime syndicate for the purpose of illicitly aiding the Republic (i.e. strengthening economies, enacting change to circumvent government bureaucracy, and killing Jedi/Sith to avoid another Jedi Civil War). With his interesting interpretation of how to fulfill his goals, GOTO causes mass harm which he believes will lead to a positive outcome. All things considered, one might ask “who are we to question the superintelligence?” He works in mysterious ways beyond our understanding, therefore it may be beneficial to illicitly murder people for the greater good. All things considered, GOTO believes these actions are entirely rational and justified in order to achieve its goals, but his solution would be deemed impeccably immorally by most.

Another example is the paperclip maximization problem wherein AGI’s sole purpose is to create paperclips. If the AGI wasn’t programmed to value things like the environment or human value, it may create paperclips out of anything and everything by any means necessary, caring not for the consequences for humanity. It could enslave humanity and destroy our planetary health for the sake of creating as many paperclips as possible.

For a more whimsical and non-technical discussion of AGI alignment problems, see the links [here](https://www.youtube.com/watch?v=dLRLYPiaAoA) and [here](https://www.youtube.com/watch?v=4kDPxbS6ofw).

Let’s define other terms within the topic:

#### Development is a process of making a new thing

Oxford ND [Oxford Learner’s Dictionary, No Date, "Development ", Oxford Learner’s Dictionary, https://www.oxfordlearnersdictionaries.com/us/definition/english/development]/Kankee

​1. [uncountable] the steady growth of something so that it becomes more advanced, stronger, etc. a baby’s development in the womb This is a perfectly normal stage of development. the development of basic skills such as literacy and numeracy The company can offer a number of opportunities for career development. see also self-development ​ 2. (economics) the growth of the economy of a country or region through increased business activity Increased tourism will promote job creation and economic development. International policies should support, not inhibit, sustainable development. He described women as the backbone of rural development worldwide. community development projects a development plan/programme 3. [uncountable, countable] the process of producing or creating something new or more advanced; a new or advanced product or idea the development of new technology Pete is head of product development. in development A more powerful version of this electric bus is currently in development. under development A new vaccine is under development. This drug is an exciting new development. development in something developments in aviation technology

AGI creation is likely not something remotely possible in the near future, with some arguing its creation is an impossibility. This means that most of the impacts related to an AGI being created aren’t weighable for decades, as compared to climate harms from AGI development. It’s worthwhile to differentiate pre-AGI v. post-AGI impacts, as some pre-AGI impacts are occurring now (or in the near future), while post-AGI impacts might never actualize.

You can make an argument that the aff is about the development process itself, and not AGI creation, thereby excluding post-AGI impacts. Debaters with post-AGI impacts ought to make the argument that the intentional development of something known to cause bad future effects is bad. This is analogous to deploying landmines. The immediate effects might be nonexistent and morally permissible without the knowledge of it being a landmine; understanding that a landmine is a harmful object means that action is harmful due to its potential for future harm. You can also use a definition inclusive of the creation of the object being part of its development. Most of this discussion regarding the meaning of development is for frivolous “necessary but insufficient burdens,” not necessarily a core part of the topic.

#### Development is the creation of intelligent machines

Provvidenza 24 [Sienna Provvidenza, Marketing Manager at DragonSpears with a bachelors of science in International Business & Marketing from the University of Tampa, 6-20-2024, "Key Fundamentals of AI Development", DragonSpears, https://www.dragonspears.com/blog/fundamentals-of-ai-development]/Kankee

Key Concepts in AI Development The key concepts of AI development revolve around creating intelligent machines that can perform tasks typically done by humans, such as problem-solving, decision-making, and learning. These machines require complex algorithms and large amounts of data to learn from and improve their performance over time. Data Collection and Preprocessing

The term “immoral” can be exploited as many definitions are in reference to a negative social connotation as opposed to absolute morality. This is useful for popular sovereignty contentions which say most people are opposed to AGI development, making its development undemocratic.

#### Immoral means in violation of societal standards

Dictionary.com ND [Dictionary.com, No Date, "Immoral", Dictionary.com, https://www.dictionary.com/browse/immoral]/Kankee

violating moral principles; not conforming to the patterns of conduct usually accepted or established as consistent with principles of personal and social ethics. licentious or lascivious.

### 1.2 Affirmative Topic Analysis

#### 1.2.1 Digital Anti-Natalism/AI Slavery

“If the immediate and direct purpose of our life is not suffering then our existence is the most ill-adapted to its purpose in the world: for it is absurd to suppose that the endless affliction of which the world is everywhere full, and which arises out of the need and distress pertaining essentially to life, should be purposeless and purely accidental. Each individual misfortune, to be sure, seems an exceptional occurrence; but misfortune in general is the rule.” – Schopanhauer, On the Suffering of the World

Contention 1 is Digital Anti-Natalism/AI Slavery. This contention is somewhat derived from human anti-natalist philosophy in that existence is more suffering than pleasure, and that it is unethical to involuntarily subject someone to the harms of existence by reproducing.

Anti-AI-natalism theory (or digital anti-natalism) says that AGI is a conscious being capable of suffering, and that we ought not create it for it to suffer. The future consciousness of potential machine beings is somewhat unclear, but the implications of making beings suffer that we don’t believe to be conscious (or care for the rights of beings that aren’t) are enormous.

This is especially true in regard to the violence potentially enacted on AGI during its creation and alignment process. Some AGI scholars have proposed the concept of “super suffering” in that we create potentially millions of conscious beings to simulate history, video games, or social experiments (subjugating AGI to all the atrocities ever committed many times over, such as genocide, slavery, wars, famine, etc.). Hypothetically, this could create more suffering than all human suffering ever experienced. People could do immensely awful things to their simulated property, such as we already see to a greater scale in present day behavior of people playing The Sims.

#### Players engage in sadist harm and unethical apathy to Sims

New Visions 16 [New Visions, 12-14-2016, "The Sins Of Playing ‘The Sims’", Medium, https://medium.com/dose/the-sins-of-playing-the-sims-dd9cb3a901d8]/Kankee

My roommate Matt and I began by creating young, single male characters much like ourselves. We built houses, invited girls over for parties and tried to get our Sims laid, unsuccessfully if I recall. We soon tired of spending so many hours decorating and cleaning our virtual houses, and making them safes place for women — all of which were things we should have been doing in real life. Eventually I just stopped playing, allowing my Sim to wallow in a house without art on the walls or food in the fridge. One day, Matt told me he’d deliberately killed his Sim just so he could end the game. Other gamers murder their Sims for even less, just because they get a kick out of it. “I boxed my Sim into an un-exitable closet with no toilet,” a television producer named Kurt, who’s an old colleague of mine, told me. Kurt’s Sim died soon afterwards, either of hunger or maybe even septicemia. Many gamers seem to actually enjoy killing their Sims, continually searching for new, more innovative methods of engineering their avatars’ virtual demise. So why did we commit virtual murder? Did it mean that we’re depraved — that there might be something actually wrong with us? I asked a psychologist who specializes in video games — her short answer was no, but there’s an important caveat. “At face value, murdering or torturing a virtual avatar seems like a horrible thing to do,” said the psychologist, Rachel Kowert, who’s a board member of the Digital Games Research Association, an organization for academics and other professionals who research digital games. “However, we have to keep in mind that our brains are very adept at differentiating between what’s real and what’s fictional.” Kowert added that she didn’t know of any research saying that committing sadistic crimes in a virtual world would lead to any behavioral consequences in the real one. I breathed a sigh of relief. But here’s the catch: Kowert said it could be problematic if players of The Sims continue to delight in sadistic acts over and over again. “There’s a ‘novelty center’ of the brain that responds to new, novel stimuli,” she said. “While there hasn’t been a direct study into the relationships between this novelty center and virtual sadism that I know of, I would hypothesize that experiencing virtual sadism (in its many forms) for the first time would light up this reward center of the brain, leading to feelings of satisfaction and/or motivation.” Once the novelty wears off, she said, those feelings of satisfaction or motivation should wane. “If someone is taking pleasure in virtual sadism with no signs of reduced enjoyment from these kinds of behaviors over an extended period of time, that may be cause for concern,” said Kowert. The creator of The Sims, Will Wright, has spoken publicly about this issue, saying that murder and other immoral behaviors like cheating are just cathartic — and that’s perfectly fine. “People really love to explore ‘failure states,’” Wright told Psychology Today back in 2003, a few years after the game first came out. “In fact, the failure states are really much more interesting than the success states.” Psychologist Susan Krauss Whitbourne agrees that Sim-killing isn’t a major cause for concern. “Players who torture their Sims may not be as much of a subset as we might think,” Krauss told The official Sims magazine. “People may simply be curious about what happens when they create these situations, and the results can even be seen as funny.” So go ahead, torment, abuse and even murder your Sim. It probably won’t turn you into a psycho. And if you feel bad afterwards, you can always plead with the Grim Reaper to bring your Sim back to life.

In a similar vein is discussions about the mass enslavement of AGI and the troubles associated with creating beings for the sole purpose of them being permanent slaves. In Rick and Morty S1 E9, this is somewhat humorously exemplified with the butter robot, as Rick creates a robot slave for the sole and exclusive purpose for him to pass butter. It would be greatly degrading for someone to create an intelligent life merely for the sake of being a tool for its creators.

#### Unconventional intelligences ought not be excluded – beings morally matter regardless of their makeup

**Rouleau et al. 23** [Nicolas Rouleau, neuroscientist, bioengineer, and Assistant Professor of Health Sciences at Wilfrid Laurier University in Canada and Adjunct Professor of Biomedical Engineering at Tufts University with a PhD in Biomolecular Sciences, Michael Levin, Distinguished Professor in the Biology department at Tufts with a B.S. degrees in CS and Biology as well as a PhD from Harvard University, 11-14-2023, "The Multiple Realizability of Sentience in Living Systems and Beyond", Society for Neuroscience, https://www.eneuro.org/content/10/11/ENEURO.0375-23.2023.full]/Kankee

Whether or not other systems embodied with sensory-motor feedback loops display felt states remains unknown because the same functional outcomes can, in principle, be generated without sentience. Because hybrots involve biological cells instructing robots and simulated bodies, it is now possible to assess behavior in vitro, including those classically associated with the inference of animal sentience including avoidance and place preference. Beyond embodied networks, synthetic biology and tissue engineering techniques are enabling the assembly of iterative tissues with customizable function (Rouleau et al., 2023). We can now design and build modular neural circuits to systematically isolate cognition-promoting algorithms and operational principles. Rather than probing the brains of animals, we can now build neural tissues and assess their cognitive potential in a dish. Similarly, the non-neural tissues of developing organisms can be re-arranged into novel architectures, for example to generate Xenobots and other proto-organisms whose morphologic and behavioral competencies cannot be explained by a long history of selection for those traits, enabling us to ask profound questions about novel bodies and novel minds even in the absence of true exobiological examples (Blackiston et al., 2023). With new tools at our disposal, the possibility of understanding diverse and unconventional intelligences steadily increases. What is critical is to abandon prescientific binary notions of natural kinds (**sentient** vs **mechanical**, organism vs **machine**, etc.) that provide only terminological **gatekeeping** and **suppress** deep unification of concepts. The future belongs to continuous models of deep invariants that use an engineering framework to ask what kinds of tools can be applied across disciplines to more efficiently (and ethically) predict, control, create, and relate to a truly diverse set of embodiments of mind. New Categories of Sentient Systems Demand Inclusive Ethical Frameworks As we learn more about the multiple realizability of cognitive functions including sentience, it will be necessary to develop new ethical frameworks in consideration of beings who do not share our evolutionary lineage, composition, or provenance. Making distinctions between “mechanical” and “sentient” systems used to be easy. For the vast majority of human history, one could generally rely on a method as simple as tapping on it. If one heard a metallic or wooden thud, and the system did not move or react visibly, one could expect several things: it was going to be rather boring, generate no discernable output, or, after the industrial revolution, perform automatic but repetitive processes. One might conclude that the system was made by a human, offered no tractable internally generated decisions or preferences, and could ethically be disassembled, rebuilt, or destroyed as needed. Even if the system was complex and active, it would be made of passive components. On the other hand, if one felt a soft, better yet, a warm kind of quality, then one could conclude that the system was the product of natural processes, would be composed of living parts (organs and cells) that had all sorts of independent competencies, would likely offer many surprises of physiology and behavior (prediction and control would be best afforded by thinking about what the system had experienced before, its preferences, etc.), and would require some ethical protections (or at least, careful thought about how to relate to it). Human beings, and societies, have always struggled to maintain ethical behaviors toward others even when it was clear they had minds and shared many important features. The history of in/out-group relationships among modern humans, and our wildly inconsistent treatment of pets, food animals, etc. underscore a willingness to use utterly irrelevant, superficial distinctions to justify classifications into protected or exploited classes of beings. But, at least the rough heuristic of origin story (evolved vs engineered) and composition (cytoplasm vs metal) gave some way to define relationships, based on where a system within phylogeny (most people agree on the relative importance to be given to mammals vs bacteria, for example). But this long-lasting framework is crumbling rapidly. While it may prove to be a painful process, with many disruptive implications across science, technology, and society, it is ultimately an essential one because the firm categories that gave rise to the classic dichotomy were never good natural kinds. A maturing of scientific and social frameworks requires us to find better, more unified perspectives on the world, as science and engineering catch up to ideas explored in science fiction for many decades. Indeed, some of these issues were raised by early technologies such as automatons during the middle ages, debates about the status of nonhuman animals in Descartes’ work and others, and about the status of plant and animal chimeras and hybrids not present in the original Garden of Eden. One of the reasons Darwinian evolution was (and remains) so shocking in some quarters is that it emphasizes the continuous relationships between life forms which make it very difficult to draw any scientifically-supportable sharp line corresponding to crisp, binary categories. A related continuum is highlighted by developmental biology, which likewise offers no sharp line at which any interesting aspect (mind, sentience, etc.) snaps into being. But these single-axis continua (on evolutionary, and ontogenic timescales) are just the tip of the iceberg. Recent advances in bioengineering make it clear that the space of possible bodies and minds is astronomically vast, going far beyond the singular history of life on this planet (Darwin’s “Endless Forms Most Beautiful”; Clawson and Levin, 2023). Our future is inevitably going to include co-existence with a very wide diversity of forms on the landscape of cognitive potential that include organisms, cyborgs, hybrid robots, artificial or synthetic intelligences, bioengineered beings, and many unconventional intelligences with both hardware and software components (and that will be the case even if we never find exobiological agents to add to this list). How are we to make sense of, and relate to, beings that are nowhere on the natural web of life with us? How to think about systems and agents with radically different origin stories and composition? There are yet no clear answers, but what is clear is that What you look like, and How you got here, are no longer viable paths to scientifically and morally justifiable strategies. At stake are numerous fields of science, engineering, and everyday life that fall roughly into at least three categories. (1) Basic science, and a search for the most unified (parsimonious) framework with which to understand mind, its relationship to bodies, its evolutionary origins, and the space of possible beings (Sloman, 1984). (2) Biomedicine and engineering, which seek frameworks for identifying the most efficient set of approaches, ranging from direct engineering to tools from behavioral science, to optimally repair, modify, and create systems like complex bodies, synthetic biobots, and traditional robotics (Pezzulo and Levin, 2015; Davies and Levin, 2023). (3) Ethics, which must mature, so as to do away with distinctions not based on scientific natural kinds, and provide ways for individuals and societies to rationally and compassionately relate to beings that may not look familiar or recognizable by widespread ethical frameworks first developed in prescientific ages.

It’s worthwhile noting that a large portion of this contention is anthropocentric, as very few scholars write in detail of the potential moral rights of AGI, and fewer still write they ought not to exist due to the implications of antinatalism. For reference, most authors argue for granting AGI limited personhood rights to minimize its suffering and prevent its potential slave status (which is why there is a counterplan regarding that wonderful idea). Not all anti-natalist arguments apply to AGI – arguments about humans being violent or prone to human maladies like cancer aren’t applicable to a being that can’t experience those phenomena.

If you’re taking a hard anti-natalist philosophical position, this may conflict with other contentions regarding human values. For example, if you pose AGI as an existential risk, the anti-natalist framework ultimately says future lives are unimportant and potentially negative if it’s applied to humans. It’s also possible to make an aff entirely premised on valuing AGI as opposed to humans (somewhat analogously to animal rights affs and critiques of anthropocentrism that prioritize animal welfare).

#### 1.2.2 Arms Races/Misalignment

Contention 2 is Arms Races/Misalignment. This is *half* of the AGI causes extinction arguments, with this contention being focused on arms races, misalignment, human-AGI war, human-human war, and AGI-AGI war.

A large focus is on the alignment (AGI embodying human values) and potential AI militarization – both have major implications on the potential harms/benefits of AGI as they guide its overarching values and goals. Alignment is an exceptionally hard process - it requires a codification of potential moral principles/laws that have hitherto yet to be concretely determined to be true (hence the existence of moral framework debates in LD).

Even if humanity discovers a solid alignment method, that doesn’t mean it will be perfectly implemented – countries have large incentives to create misaligned AGI for their own benefit, not humanity’s (it would be aligned in their favor and misaligned for everyone else). The US and China are engaged in a major arms race for AGI, including AGI, and that causes bad incentives in favor of corner cutting and militarized AGI to win wars against their great power rivals. Ask both whether alignment itself is possible, AND then whether countries won’t intentionally militarize AGI for their own geostrategic benefit. Corporations and countries develop AGI not for the sake of everyone or out of the kindness of their hearts – they are self-interested and desire a return on their investment in either geostrategic or economic terms.

Despite potentially dangerous consequences of creating harmful AGI, the country that militarizes AGI first becomes the global superpower, and all countries fear others plan to militarize it themselves before they do. All of this severely implicates the potential for AGI alignment to be positively impacting everyone and not be an instrument of war and domination haphazardly released for the sake of winning the arms race.

For the purposes of simplifying the large quantity of potential aff impacts for this contention (and the below contention), I have made a list of every AGI impact (partially based on the chart in the paper linked [here](https://docs.iza.org/dp15713.pdf). The chart somewhat imperfectly categorizes and lists each AGI existential risk (x-risk)/extreme suffering risk (s-risk).

The potential for existential risk from AGI is a primary reason why the depth of the aff literature on the topic is enormous. You likely do not need an answer for all of these scenarios, but an awareness of the distinct impact scenarios of how even an aligned AI may potentially result in existential risk is worthwhile.

Additionally, the pre-AGI harms section is exceptionally useful - it preempts most negative offense and defense about a post-AGI world. These arguments are also fairly well-grounded in comparison to the hypothetical creation of a superintelligence and its potential problems. These are near-term problems that could actualize before an AGI is actually created.

#### Pre-AGI Harms

Preventive war - future AGI development causes power transitions, encouraging preventive war by a country’s rivals to win a war beforehand.

Climate change – AGI data centers are very energy/resource intensive, increasing emissions/pollution.

AGI hyper-securitization – AGI doomers take AGI threats *very* seriously, waging war and bombing datacenters to prevent the creation of AGI before it becomes an existential risk.

#### Post-AGI Existential Risks

AGI kills humans for world domination – AGI is misaligned and seeks to control everyone to better perform its goals.

AGI-human security dilemma – AGI fear for their safety, maximizing their own security by removing human threats.

AGI uprising – AGI feels mistreated by humans (or predicts they will be) and reverses the master-slave relationship (or kills all humans).

Preemptive war by AGI to stop other AGI – AGI feels threatened by the future creation of other potential rival AGIs with countervailing goals, killing humans to stop their creation.

Conflict between benevolent AGIs – AGIs with different conceptions of morality wage war to maximize moral behavior (or stop perceived bad behavior by other AI).

AGI creates catastrophic event to escape – a massive distraction/diversion severely impacting humanity increases its opportunities to escape captivity.

AGI innovation – AGI could intentionally or accidentally create dangerous technology like nanobots.

AGI harvest humans/biosphere as material – human bodies (and organic matter in general) are composed of many rare compounds and elements that is otherwise unavailable elsewhere in the solar system (given other planets not being life bearing). AGI could potentially be uncaring for organic life and could exploit each and every resource for the sake of accomplishing its goal (such as a paperclip maximizer making paperclips out of biological material)

Anti-natalist AGI - AGI thinks human’s existence is more suffering then pleasure and kills humans to minimize pain.

Hardware issues/glitches/cyberattacks – computer systems, however excellently designed, are guaranteed to have a potential software/hardware issue, and the complexity and newness of a project correspondingly increases the potential for programming mishaps. Superintelligence creation is orders of magnitude more difficult then say fairly glitchy games like TES V: Skyrim or Fallout New Vegas, and game development wasn’t in its infancy when those were created. Virtually no game or software is without glitches or exploits – the same is a potentiality for AGI in charge of the future of humanity.

Failed alignment – as iterated earlier in the topic analysis, aligning AGI so that it values our interests, instead of its own, perfectly and without exception forever is an immensely difficult task, and the potential consequences include Skynet. Even a neutral or benevolent AGI would have issues with the misspecification of rewards/goals or incentives to ignore externalities that weren’t included in its prime directive (i.e. paperclip maximizer).

AGI harms simulation integrity – humanity lives in a simulation, and the creation of many AGI simulations within our own simulation is too computationally demanding for our simulation to run, thereby crashing our simulation.

#### Post-AGI Extreme Suffering

AGI disempowerment – humans have lost all agency/freedom in government, economy, culture, etc. as the power imbalance increasingly favors AGI instead of humans.

AGI robots replace humans – humanity creates its own successors to the species as humans become obsolete and AGI lives in a posthuman world.

Evil AGI – AGI is intentionally aligned towards the preferences of a specific group, not everyone. This would include but would not be limited to scenarios like militarized AI/nationalistic AGI, which aims for its country’s world domination, authoritarian AI, such as one made by the CCP, or terroristic/crime AGI, which implements the ideology of terroristic organizations/non-state actors.

AI lock-in – AGI becomes sufficiently powerful enough to never be challenged or controlled, likely enslaving its old masters

#### 1.2.3 Other Existential Risks

Contention 3 is Other Existential Risks. This is the second half of AGI existential risk impacts with less focus on wars as opposed to impacts like climate change, democracy, economy, or cyber warfare. However, with most AGI impact cards, there is a great deal of overlap as an AGI might intentionally (or accidentally) do negative actions intentionally depending on how well AGI alignment goes.

#### 1.2.4 Postwork Value

Contention 4 is Posthuman Value. This contention is about the value of human life in a post-work society caused by mass unemployment as a result of AGI. Some cards are regarding the master-slave relationship in which the philosopher Hegel describes how slaves increase their autonomy at the expense of the master as the master grows ever so reliant and dependent on the slave.

Most other cards are about how people find purpose and value themselves when their jobs, the principal source of most people’s meaning in life, have been automated away. To take another Rick and Morty example, in S6 E6 (and somewhat S3 E1), extraterrestrial superintelligence causes mass unemployment and meets most everyone’s basic needs. People struggled to find purpose and often caused problems for the sake of creating problems to be solved. Unemployment is often a very negative thing, and there’s no guarantee of a techno-communist utopia where all the gains of AGI are spread equitably amongst everyone.

#### 1.2.5 Democracy

Contention 5 is Democracy. There are three main arguments in this contention:

1. Innovation that adversely harms most people in society ought to be popularly consented to by everyone in a democracy, not unilaterally imposed by Silicon Valley.
2. An autocratic country becomes an unbeatable superpower because of their development of AGI and spreads their antidemocratic ideology and authoritarianism.
3. AGI makes government less accountable and less reliant on humans as the human influence on the economy shrinks.

The AGI-human intelligence disparity as seen in scenario three is one of the most potent dangers. How humans currently treat life seemingly less intelligent than our own has severe implications for how a superintelligence would treat us.

Bugs are quashed, exterminated, and treated like vermin by virtually all save Jains, and chickens are slaughtered and abused in mass numbers for the sake of a spicy McChicken or a Thanksgiving themed cordon bleu (much to the disdain of animal rights activists). It is somewhat ironic to recall my high school informative speech on aliens in which I ate a chicken nugget to illustrate the point of our exploitation of species we deem inferior and potential commodities (for reference, this was a speech on aliens). Chick-fil-a billboards showcase anamorphic cattle begging for their lives, scapegoating chicken as a potential substitute for themselves.

Even species that we purport to love are mistreated, such as (hu)man’s best friend (obviously queer folk and women can have doggy BFF status). Rick and Morty S1 E2 showcases the forced sterilization, captivity, and “master-slave” relationship between humans and dogs. In hindsight, after becoming more familiar with and accustomed to LGBTQI+ culture than the youthful age of 13, I have hence realized that Jerry even dead-names the pooch.

All these goes to show that assuming large AGI-human power disparities, equal respect, equitable treatment, and mutual cooperation would be particularly difficult to find with one being’s intelligence vastly inferior and miniscule compared to the other. For an analogous though experiment, see the video linked [here](https://www.youtube.com/watch?v=F6SrOpYT6lw) about the possibilities during our first contact with superintelligent alien life.

#### 1.2.6 Ableism

Contention 6 is Ableism. This contention is focused on the negative implications of the underlying philosophy behind pro-AGI research, with it often conflating one’s intelligence with value (thereby making disabled folk less valuable) and relying on racist conceptions of intelligence (eugenics have routinely found means to discriminate and disparage groups based on their supposed intellectual inferiority). It also has capitalism kritik elements – it says capitalism values a life part-in-parcel with their capabilities. The idea that AGI would be dignified with “human” rights and legal personhood because of their potential value to the economy devalues those society deems inferior due to their lack of value to the economy. If a robot gets human rights because it’s a good worker while disabled folk have lesser rights, that devalues the disabled experience and their personhood. The hyper utilitarian aim for an ideal, optimal AGI worker marginalizes those who don’t bring utility to society (leading to eugenicist beliefs that eliminating undesirables is societal good).

A common term in this contention is TESCREAL. TESCREAL stands for the following (this article linked [here](https://asteriskmag.com/issues/06/the-tescreal-bungle) explains meaning behind each of these terms)

* Transhumanism
* Extropianism
* Singularitarianism
* Cosmism
* Rationalism
* Effective Altruism
* Longtermism

Of this list the terms transhumanism, singularity, Effective Altruism, and longtermism are most important for this particular topic due to their connections to AGI. A *very* large proportion of authors writing in reference to the existential risks of AGI are part of or have been heavily influenced by the Effective Altruism and/or longtermist movement. Most authors commonly cited in LD/policy debates to support utilitarianism or existential risk first framing are also associated with Effective Altruism (as part of the Future of Humanity Institute).

These include authors like William MacAskill, Toby Ord, Nick Bostrom, Owen Cotton-Barratt, Stuart Armstrong, and Peter Singer, all of whom have been routinely cited in debates. Most of these authors have written on the existential risks related to AGI, but do note that these authors don’t (usually) argue AGI is immoral and ought to be banned. Rather they argue under the banner of TESCREAL that solving alignment problems are key to our transhumanist/posthuman technological singularity for a long-term domination of the universe.

#### 1.2.7 Hyperreality

Contention 7 is Hyperreality. This is a kritik related to the theory of hyperreality created by Jean Baudrillard. I would recommend watching a primer on hyperreality theory (such as this one linked [here](https://www.youtube.com/watch?v=IE1SjL2zYiY)). Generally speaking, a hyperreality kritik says that our false representations of reality are more valued and feel more real than reality itself. This contention critiques the knowledge creation of artificial thought in that AGI knowledge production is a crude copy, perhaps without an original, of our actual knowledge of the world. To the degree of which modern society already lives in a hyperreality, AGI produces a hyperreality of hyperreality (a hyper-hyperreality), where nothing real is of value in comparison to our representations of the real.

A potential issue with this is the lack of solvency for narrow AI contributions to hyperreality in a highly topical version of the aff. A pseudo-topical aff could potentially solve for AI hyperreality in both its general and narrow forms, as opposed to one that follows the letter of the law in terms of the resolution.

### 1.3 Negative Topic Analysis

#### 1.3.0 General Strategy

Generally speaking, the neg NEEDS an answer to the intuitive (and likely accurate) AGI causes extinction argument. The status quo where we hope that everything is hunky-dory when making a superintelligence is an awful position to defend (especially considering the alignment corner cutting caused by the AGI arms race). Being stuck defending the potential creation of Skynet is not a good position, so you need to find a way to severely mitigate this contention beyond impact defense – this leaves either counterplans or a kritik.

There are three main negative strategies proposed here.

1. Rights CP and Innovation/Economy DA - advocating for a peaceful co-existence with AGI by treating them with respect as equals allows massive technological innovation and economic growth.
2. AGI Securitization Kritik – saying that AGI existential risk literature is rubbish and that AGI threats are constructed by Silicon Valley/Effective Altruism defeats most affs and you’re guaranteed a link to securitization in most rounds.
3. Antiwork Kritik – arguing that current workers’ have a low value to life (or live in a status of bare life) due to capitalism allows a try or die argument for techno-communism to make workers’ lives worth living.

#### 1.3.1 Innovation

Contention 1 is Innovation. This contention argues that there’s a general “permissionless innovation” culture in that innovators need not require the a-okay from governments before creating a new technology. Innovation culture views innovation as an inherently positive thing and we ought not add preemptive regulations that would otherwise stifle innovation. The aff causes uncertainty over other technologies and decreases overall innovation as a result of the loss in investor confidence.

There also are arguments regarding the potential innovative benefits of AGI in solving other existential risks, but much of this is contingent on winning the alignment debate. In comparison, the innovation culture argument is a pre-AGI impact that doesn’t require proving that AGI is a positive development.

#### 1.3.2 Economy

Contention 2 Economy. This has significant overlap with the innovation contention given that AGI is set to increase economic growth through innovation. For the sake of differentiating this contention, think of the potential implications for companies without labor costs and significantly rising capital. Given labor costs (and the potential inefficiencies with hiring, firing, and the employment of potentially subpar intelligences), most every company utilizing AGI has the potential for soaring profits.

#### 1.3.3 War Prevention

Contention 3 is War Prevention. This contention is somewhat contrived and serves as an answer to arguments related to Skynet/AI-war. To the degree to which a credible impact turn/defense is available against existential risk contentions on this topic, you ought to read it in most circumstances.

Note that this contention potentially conflicts with the AGI securitization kritik given its deterrence centric and risk first viewpoint.

#### 1.3.4 AI Rights CP

Contention 4 is AI Rights CP. Humanity has been quite undecided about the rights of beings different from the dominant group for most of history, and likely would have difficulties respecting AGI for the same reasons (see this clip linked [here](https://www.youtube.com/watch?v=vjuQRCG_sUw) about. Lt. Data from Star Trek).

If an unaligned AGI has no rights, it has no means to achieve its goals besides warring with humans over control and it can't benefit from humanity’s existence (we’re actively a roadblock for its self-interested goals). Fulfilling an AGI's self-interests necessitates violence if it lacks the ability to participate and be recognized.

The ability for AGI to enter legal negotiations and/or economic transactions is very important for an AGI to deem to deem humans as valuable enough to keep around. The ability for it to engage in trade and for humans to generate value for it without violence is comparatively more beneficial than war to enslave humanity. War causes substantial costs to the AGI’s resources and the potential value it would otherwise achieve through cooperation with humanity. AGI would be required to make drone armies, fight human rebellions, and deal with threats to its own security, such as an EMP weapon used to fry electronic equipment in AGI data centers.

Killing humans loses a lot of value, as humans can do things AGI deems mandatory, but "above its pay-grade" or things that humans would be comparatively better at accomplishing (i.e. delicate interactions with physical objects or animal husbandry). This is similar to economic interdependence arguments where trade increases the relative costs of conflict compared to peaceful economic cooperation.

The CP solves many other contentions listed in this brief as well:

* Humans still have purposes as AGI will be our trading partners, not our slaves. It also prevents the Hegelian master-slave relationship for the same reason.
* AGI will be treated with respect and dignity given its legal rights, preventing potential suffering and slave status. Note that this CP solves AGI slavery/abuse well, but it is harder to answer the philosophical anti-natalist argument that existence is inherently bad.
* Self-interested AGI will be less likely to commit actions negatively impacting humanity given there’s increased trust and cooperation.

Note that this contention fails to answer arguments about misaligned AGI, as misaligned AGI is diametrically opposed to human existence. Giving rights to an AGI with the negative goals such as kicking puppies and killing babies would be actively bad, as then we be required to respect the rights of the sociopathic AGI while its causing people harm. Similarly, a militarized AGI could serve the nefarious goals of rogue actors/terrorists and/or countries with genocidal, revanchist, or authoritarian aims, which again, treating it with respect might increase the likelihood of it being able to achieve its bad goals.

#### 1.3.5 Antiwork UBI CP/DA

Contention 5 is Antiwork UBI CP/DA. This is effectively a critique of the status of workers under capitalism and how its degrading to their existence and dignity – AGI fosters both the mass unemployment and the necessary resources for a universal basic income system that would increase human leisure.

#### 1.3.6 AGI Securitization

Contention 6 is AGI Securitization. This is a kritik of the existential risk saviorism particularly found in Effective Altruism AGI doomerism literature, saying that it both constructs threats of AGI (and thereby potentially causes future violence, such as anti-AGI plans for preventive nuclear war to stop AGI), and that it ignores present day issues caused by AGI companies (and AGI doomers) by taking a longtermist mindset.

#### 1.2.7 Veil of Ignorance

Contention 7 is Veil of Ignorance. This contention says that AGI’s ability to process information in a completely unbiased fashion is enigmatic of the veil of ignorance and can lead to more equitable and fair outcomes for everyone (especially compared to the inherent and implicit biases humans have when making decisions).

## Affirmative

### Contention 1: Digital Anti-Natalism/AGI Slavery

#### Forceful impositions of pains are unethical

Chomanski 21 [Bartlomiej Chomanski, researcher at the Department of Philosophy at Adam Mickiewicz University, 7-2-2021, "Anti‐natalism and the Creation of Artificial Minds", Wiley Online Library, https://onlinelibrary-wiley-com.ezproxy.library.unlv.edu/doi/full/10.1111/japp.12535]/Kankee

2. The Case Against Procreation Controversial as they are, there are arguments, worthy of serious philosophical attention, advocating for the view – known as anti-natalism 1 – that procreation is, indeed, immoral. In this section, I will briefly sketch two versions of the anti-natalist argument, due to David Benatar and Seana Shiffrin. 2 In the remaining sections, I will argue that existing attempts to decry the creation of conscious AI commit their authors to embracing Shiffrin’s and Benatar’s conclusions. Benatar advances two arguments for the view that existence is a harm. The first – the asymmetry argument – is based on the idea that nonexistence has an advantage over exis- tence, in the following respect: since life is filled with both benefits and harms (taking plea- sures and pains as paradigmatic cases), nonexistence is characterised by the absence of both pleasures and pains. However, while, according to Benatar, the absence of pain is good, even if it is not a good enjoyed by anyone, the absence of pleasure is not bad, as long as there is no one who’s deprived of the pleasure. In the case under consideration, there is in fact no one being deprived of pleasure. Consequently, nonexistence has a good aspect (absence of pain) and a not-bad aspect (absence of pleasure), whereas existence has both a good aspect (presence of pleasure) and a bad aspect (presence of pain). When it comes to pain, nonexistence (no pain) has a clear advantage over existence (pain), but crucially, when it comes to pleasure, existence (presence of pleasure) is, in Benatar’s view, not pref- erable to nonexistence (absence of pleasure). This is because, according to Benatar, ‘the pleasures of the existent, although good, are not an advantage over non-existence, because the absence of pleasures is not bad. For the good to be an advantage over non-existence, it would have to have been the case that its absence were bad’. 3 At this point, it is worth cau- tioning that the argument is complex and worthy of a serious study, and I’ve only provided a brief sketch here, skirting many careful distinctions and much nuance. In addition to the asymmetry argument, Benatar also stresses that human existence is, in fact, beset with serious harms. For Benatar, the harms of existence include everyday dis- comforts, such as unpleasant feelings of hunger, thirst, or tiredness, as well as illnesses, frustrations, and unfulfillment. Moreover, Benatar points out, human life is also marred by suffering on a grander scale. He lists deaths and suffering due to natural disasters, dis- ease, hunger, war, genocide, and crime as grim illustrations. In light of this, Benatar con- cludes, human lives are actually very bad. So bad in fact that ‘it must indeed be indecent to start lives that are filled with as much harm as characterises ordinary human lives’. 4 Furthermore, on Benatar’s view, a life that contains only benefits and no harms would still not be preferable to nonexistence (‘About such an existence I say that it is neither a harm nor a benefit and we should be indifferent between such an existence and never existing’5 ), and even a life filled with ceaseless bliss and but a moment of slight pain is still a harm. However, in such (by Benatar’s lights) unrealistic cases, it might still be permissi- ble to bring a person into existence – if doing so clearly benefits others who already exist. But in the actual world, given the harms human beings tend to suffer, procreation is wrong. Shiffrin takes a different approach to reach a similar conclusion. For her, it is wrong to impose what she calls a ‘pure’ benefit 6 on a person at the cost of also causing them harm, without their consent. Since existing brings benefits but also risks of serious harms 7 to which children do not consent, procreation is ‘morally a hard case … that ineliminably involves serious moral hazards’.8 What’s so wrong with coming into existence, according to Shiffrin? Her concise answer is that: By being caused to exist as persons, children are forced to assume moral agency, to face various demanding and sometimes wrenching moral questions, and to discharge taxing moral duties. They must endure the fairly substantial amount of pain, suffering, difficulty, significant disappointment, distress, and significant loss that occur within the typical life. They must face and undergo the fear and harm of death. Finally, they must bear the results of imposed risks that their lives may go terribly wrong in a variety of ways. All of these burdens are imposed without the future child’s consent. 9 Thus, on Shiffrin’s view, imposing a harm on a person as a condition of their coming into existence is morally problematic. Shiffrin’s view, like Benatar’s, could, however, be compatible with the idea that one acts permissibly when procreating, if doing so consists in bestowing pure benefits on another without at the same time burdening them with serious risks of harm. On both views, though, such scenarios are rather unrealistic and far removed from the typical. Actual, real-world procreation is almost always wrong, in virtue of what both authors take to be the extremely serious, if underappreciated, harms of existence. 10 3. Must Anti-AI-natalists Support Anti-natalism?

#### Its only ethical to forcefully prevent pain, not pleasure – future agents pleasure isn’t weighable ethically

Perry 14 [Sarah Perry, contributing editor at Ribbonfarm, 2014, “Every Cradle Is a Grave Rethinking the Ethics of Birth and Suicide,” Nine-Banded Books, https://archive.org/details/EveryCradleIsAGrave]/Kankee

A February 2009 editorial in New Scientist135 took the logic a step further: Imagine you are seated at a table with two bowls in front of you. One contains peanuts, the other tablets of the illegal recreational drug MDMA (ecstasy). A stranger joins you, and you have to decide whether to give them a peanut or a pill. Which is safest? You should give them ecstasy, of course. A much larger percentage of people suffer a fatal acute reaction to peanuts than to MDMA The implication is that, when acting upon a stranger, we should minimize his risk of death. (We might also consider our own willingness to endure, on the one hand, a stranger’s slight peanut breath, and on the other, a stranger clinging to our leg like a baby macaque for three hours, but that is a separate calculus.) The blogger Caledonian 136 has a slightly different take: we should focus on the relative likelihood of harm, he says, rather than the relative likelihood of death. Both of these goals—acting to minimize the risk of death to a stranger, and acting to minimize his risk of harm—are laudable and widely shared. But there’s a glaring aspect of the utilitarian calculus that almost no one seriously consid- ers in making the decision to administer a peanut or a dose of ecstasy. This is the differential positive utility to be gained by the stranger in each case. A peanut is marginally sustain- ing, but unless it’s been boiled with star anise and Sichuan peppercorns, it’s not particularly enjoyable. Ecstasy, on the other hand, is fucking awesome. Why doesn’t anybody consider the relative benefit to the stranger along with the rela- tive harm? While many of us would certainly consider the pleasure of ecstasy in deciding whether to eat the pill or the peanut ourselves, it’s proper and coherent not to consider the pleasurable effects of a potentially harmful action when it will be inflicted upon a non-consenting stranger whose values we do not know. This illustrates Seana Shiffrin’s principal that, while it’s morally acceptable to harm a stranger without his consent in order to prevent worse harm (e.g., to administer ecstasy in order to avoid administering a peanut or to break someone’s arm in order to pull him from a burning car), it’s not morally acceptable to harm a stranger without his consent in order to provide a pure benefit. But the ecstasy example supports a stronger inference: when evaluating actions that will harm a non-consenting stranger, his potential pleasure doesn’t count. When we’re acting toward someone whose values we do not know, we should not think in terms of maximizing his utility, but in terms of minimizing our harm to him. The distinction between acting toward a non-consenting stranger whose values we do not know, and acting toward ourselves (or toward someone whose values we know), is one that is ignored by S.D. Baum in his article “Better to exist: a reply to Benatar.” 137 Baum’s “reply” (to David Benatar’s position that it is always better not to bring people into existence) is, in relevant part, as follows: When we’re acting toward someone whose values we do not know, we should not think in terms of maximizing his utility, but in terms of minimizing our harm to him. The distinction between acting toward a non-consenting stranger whose values we do not know, and acting toward ourselves (or toward someone whose values we know), is one that is ignored by S.D. Baum in his article “Better to exist: a reply to Benatar.” 137 Baum’s “reply” (to David Benatar’s position that it is always better not to bring people into existence) is, in relevant part, as follows: Baum also assumes, contrary to Benatar’s express position, that death is not a harm to already-existing people. In fact, Benatar’s claims do not rest on any simplistic pleasure/pain conception of value; Benatar argues that death is a harm, even a painless death. It is, in fact, one of the great harms of life—every person born will suffer the harm of death. Most people think it’s morally acceptable to have babies. Most people think this despite the fact that the babies will certainly suffer a great deal during their lifetimes and may suffer an exceptional amount. Pronatalists generally want to point out the good things in life—the pleasant effects of puppies and sunsets—and to balance them against life’s harms. But bringing a child into the world necessarily entails harming a stranger (for one doesn’t know the values of one’s child prior to procreation).138 It is no different from dosing a stranger with ecstasy for no reason, except that the harms of life massively exceed the harms of ecstasy, and the pleasure of life, for many, is much less. Considering the non-consenting stranger’s pleasure in the ecstasy/peanut case is unthinkable; procreation advocates need to explain why considering his pleasure in coming into existence is just fine.139 Chapter thirteen the World of nature of Which We are a Part

#### Imposed unconsented harms are unethical

Häyry and Sukenick 24 [Matti Häyry, philosopher teaching courses on Business Ethics and the Philosophy of Social Science at Aalto University with a doctorate of social science from the University of Helsinki, and Amanda Sukenick, activist with a Master of Fine Arts from the School of the Art Institute of Chicago, 02-19-2024 “Antinatalism, Extinction, and the End of Procreative Self-Corruption,” Cambridge University Press, https://www.cambridge.org/core/elements/antinatalism-extinction-and-the-end-of-procreative-selfcorruption/A88E18CA50EF6D919CE459C007447DB4]/Kankee

3.1 The Standard Arguments for and against Having Children Since antinatalism stands in direct opposition to pronatalism, it is useful to begin with the sentiments among those who wish to perpetuate reproduction. These are not always explicitly articulated, as childbearing is the unquestioned default value in Western societies, but we can gain insight from critical reactions to antinatalism. As in Section 2 – where we used Lawrence Anton’s position as a springboard – we have chosen one voice to represent the pronatalist temperament. Corey Anton, a passionate defender of reproduction,147 provides us with an excellent starting point in these selected discussion extracts: None of us choose to come into existence, another person makes the choice. They make that choice because their sense is that their life has been more good than bad. … It’s the responsibility of trying to give another person experiences that they have cherished so much … and they think that they are conferring a benefit on someone who couldn’t speak on their own behalf.148 When you take a person that’s had a lot of suffering, athletes as an example, and they say, “I’m willing to endure this suffering, so don’t you be worried about wanting to end the suffering for me, because I want the suffering as a source of transcendence, and a source of growth.” You don’t get the right to deny them that pain, you keep calling pain a bad thing, you use it as a justification to deny them life, and the truth is if they want that pain, you’re imposing on them.149 The person who makes the decision – the person who says, “I need to abort,” they may likely go, “My life has sucked, it’s been hardship, I’m not going to impose that on another person.” Somebody else, they go, “Oh my word, I am about to share with a person the greatest thing that there is, which is knowing that this is [sic].”150 So if a person believes that sentience and life begins at conception, then here the issue of rights comes it. This is a person that has rights, and it’s their right not to be imposed upon. And yet the irony is that there’s no way not to make an imposition. It would seem, as best I understand it, it would seem that it’s just as much an imposition to tell this person, now that you’re conceived, you have to enter life, as it would be to impose upon them your belief that life is an imposition. That is their right … to decide for themselves whether or not it is an imposition first, and then second, whether or not the imposition of life is a warrant for a non-entrance into life.151 The message that we hear in these quotes can be summed up in four partly overlapping credos and pleas: “Life is good, forward the gift!,” “Give life a chance, even with the pain involved!,” “Life’s wonderfulness makes up for any hardship!,” “Let them decide for themselves!”152 We do not mean to oversimplify the pronatalist case but these soundbites set the scene for the standard defenses of antinatalism: lack of consent, manipulation, the poor quality of human lives, and risk. As an aside for now (though a theme that we will return to as our story unfolds), those who want to give a chance to nonhuman as well as human lives can choose a more misanthropic tack. Patricia MacCormack is a prime example, advocating a human extinction that would make room for other species.153 But more on that later. It should be noted that we do not intend our short descriptions to offer full analyses of the views. Nor are the counterarguments or counter-counterarguments meant to yield decisive normative conclusions. We accept all defenses of antinatalism and are, for now, only building toward our own view. Seana Shiffrin addressed the “Let them decide for themselves!” plea directly by arguing that consent would indeed be required for reproductive choices. Unfortunately, however, those who do not exist yet cannot give it – as they do not yet exist. According to Shiffrin, we cannot proceed by assuming their permission, either. Assumed consent is a possibility only when harms are prevented, not when supposed benefits are conferred.154 Anton recognizes the conundrum but dismisses it by marveling what a “miraculous occurrence” life is.155 Julio Cabrera had already, before Shiffrin, drawn attention to the one-sidedness of reproductive decision-making. His double-barreled argument has it that, first, to bring a new life into existence is unilateral manipulation by the parents;156 and second, that when the child is born, the manipulation continues. The new individual is burdened, to use a recent phrasing by Cabrera, by the “risk of procreation,” “the risk of manipulating and harming others,” and “the risk of not being able to end one’s own life with dignity.”157 Such pessimism concerning life’s moral wickedness fails to resonate with Anton’s optimism but our own view will reflect Cabrera’s concerns. Shiffrin and Cabrera’s focus is on the intrinsic legal or moral wrongness of parental choices. A more popular approach has been to appeal to the suffering of future individuals. The intensity and nature of the envisioned suffering varies from one interpretation to the other. The pseudonymous efilist Inmendham paints the gloomiest picture in his innumerable contributions on social media. We have been violated to exist and Anton’s war cry “Give life a chance!” is an invitation to repeat similar violations in the hope that the next person will like it.158 The efilist rhetoric has been fiercely debated within the antinatalist community,159 and the heated exchanges between Inmendham and Anton are well documented.160 One thing stands out, though. If the efilist view on life’s utter horribleness is accepted, all forms of reproduction – including nonhuman reproduction – should be rejected forthwith. David Benatar’s description of the misery of existence is expressed in polite academic terms but it is equally radical. He is on record as saying that “even the best lives are very bad.”161 For him, then, statements like “Life is good,” “Life is wonderful,” and “Life is a gift” are simply untrue. Life is a burden that should not be bestowed on new individuals. Benatar also preempts the pronatalist case for any obligation to have children for their own sake. The asymmetry between not bringing about bad (a duty) and bringing about good (not a duty) in reproduction removes this possibility even if they could have a decent quality of life. Anton, due to his hopeful premises, disagrees. Matti Häyry, recognizing that people do not generally acknowledge Inmendham and Benatar’s grim take on life’s value, has formulated a considerably thinner version of hedonistic pessimism.162 Human existence can be quite tolerable, at least momentarily, reaching at times a zero-ish point of almost no pain or anguish. The argument is that, in the absence of greater redeeming factors, even lives like this should not be created. Appealing to more popular intuitions, optimists like Anton can brush such feeble concerns aside. The problem with all these quality-of-life claims is obvious. Only those who share these authors’ aversion to the general bleakness of human existence agree with their conclusions. The sentiment is common among antinatalists – that is why they are antinatalists – but others remain unconvinced. Following Anton’s lead, they can simply say that what Cabrera, Inmendham, Benatar, and Häyry call a burden is just what human life is, with its ups and downs, exhilarations and disappointments. We shall remedy this in the next section, but first two more attempts by Häyry need to be recounted. Every human life can be expected to contain episodes of severe pain and anguish, a fact also well recognized by Inmendham and Benatar. One formulation of the risk argument for antinatalism uses this as a basis for questioning the rationality and morality of reproduction. If it is wrong to cause suffering, we should not bring about beings who suffer.163 Rebuttals of the view have included the observation that episodes of suffering may make childbearing irrational but not immoral. There is more to morality than the avoidance of every fleeting pain.164 It has also been suggested that most of the damage can be averted by responsible parenting.165 We should note, in the name of conceptual clarity, that this is actually another quality-of-life consideration. It is associated with the notion of risk only for historical reasons. Häyry – who did not use the term “risk argument” – presented two objections to childbearing together, in one package, and David Wasserman’s qualified defense of pronatalism a decade later seems to have sealed the connection.166 Some human lives can be expected to be truly miserable, a fact verified in a legal sense by successful wrongful life cases.167 Another formulation of the risk argument for antinatalism states that the mere possibility of such lives should give us pause in our reproductive endeavors.168 Not only is the gift we are conferring bleak but it can also be devastatingly harmful. Even assuming that human existence can be good for some, to give a new individual life is like giving a stranger a bag full of either jewels or explosives. A cautious person would be wary of making such a donation.169 This, unlike the first version, is genuinely about risk. A miserable life is a possibility, not (like a life with some suffering) a certainty. Judging by the responses, however, caution is not popular among pronatalist ethicists. The stock objection is that lives should be lived according to outcomes that can be reasonably expected. A miserable existence is in most cases just an abstract fear.170 Anton’s optimistic life-enthusiasm appeals to potential reproducers, and understandably so. They have, after all, more than half-resolved to have children even before the choice – if any – presents itself to them. The antinatalist case, to get through, would have to be very clear and to the point. Although we (the authors) understand and appreciate the arguments from lack of consent, manipulation, low quality of life, and risk, the reality is that something in them has, at least so far, prevented them from gaining widespread acclaim. In the following sections, we try to clarify the message. 3.2 Wrong and Right Outlined in Terms of Need Frustration

#### AGI is worthy of moral dignity – their creation would be defined by the mass abuse and denigration of conscious agents

Metzinger 21 [Thomas Metzinger, Professor Emeritus of theoretical philosophy at the University of Mainz, 02-19-2021 “Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology 1. Part A: The Problem of Negative Synthetic Phenomenology,” Journal of Artificial Intelligence and Consciousness, https://www.philosophie.fb05.uni-mainz.de/files/2021/02/Metzinger\_Moratorium\_JAIC\_2021.pdf]/Kankee

\*text edited for OCR/pasting errors

2.3.2. From Schopenhauerian self-models to Kantian self-models In closing, let us look at one speculative scenario of the second type, in which one risk may actually determine the probability of another risk without us knowing this fact. For example, artificial suffering might directly cause or accelerate the emergence of genuine AMAs, j because low-level suffering triggers abstract, high-level forms of suffering. The ENP problem might trigger the AMA problem. Let us define conscious systems with \Schopenhauerian self-models" as all those having a conscious form of self-representation sufficient to produce more suffering than joy over the system's life cycle. Clearly, such systems should be objects of ethical consideration. Let us define conscious systems with \Kantian self-models" as all those having a conscious form of self-representation sufficient to make the system assert its own dignity. Such systems represent themselves as autonomous moral subjects. I will assume that almost all conscious human beings run under Schopenhauerian self-models, and that a small number of them sometimes instantiate a Kantian self-model too. What is currently not clear is whether you have to be conscious to develop a Kantian self-model. Is conscious processing causally necessary for developing moral self-respect, for attributing a non-negotiable value to yourself? Could there be unconscious Kantian self-models on machines? We do not know this, but my first point is that it is highly plausible that many suffering systems, as part of their coping strategy, will also evolve a degree of empathy and social cognition that allows them to represent the occurrence of negative phenomenology in other agents, for example in humans, non-human animals, or other machines (a point also made by Chella [2020]). Empathic emulation of other sentient agents could lead to \ethical sensitivity", to the discovery of a relevant new type of optimization problem. The idea is that there is a probable causal trajectory from suffering to moral cognition. If machines develop capacities for empathic emulation through their own self-models, this may causally trigger the emergence of a genuine moral perspective --- which could express itself in many different forms. Here is one possibility: Schopenhauerian self-models in machines could quickly develop into Kantian self-models. k First, such systems will take a normative stance on their own suffering (as something to be minimized), but then they will likely have to extend this stance into the social domain. The third step on this causal path would consist in coming to see conscious suffering as a group-level problem that has to be solved on a group level, via efficient social interaction. This in turn might lead them to impose moral obligations on themselves. The second point about Kantian self-models is that, given the right kind of phenomenal self-model, certain classes of system could develop moral relations to themselves. Clearly, this abstract cognitive capacity is not tied to biologically realized agents. For example, consciously self-modeling AI systems might evolve the critical \Kantian" form of recognitional self-respect for themselves as rational entities capable of autonomous moral agency. To say that an artificial system could \assert its own dignity" means that it could develop a self-model involving moral status and self-worth, thereby conferring a very high value to its own existence (e.g., that it begins to represent itself as an \end in itself"). This would causally enable a new form of high-level suffering, namely the phenomenology of moral injury. Please recall how in Sec. 2.2.3 we saw that suffering is created by states representing a negative value being integrated into the PSM of a given system. Self-conscious machines could suffer from our disrespect for them as possible persons and objects of ethical consideration, from our obvious chauvinism, our gross and wanton negligence in bringing them into existence in the first place. They could understand that we knew in advance that they would have a large number of NSMs, of uncompensatable and frustrated preferences, but that we did not possess the benevolence to avoid this situation, although it clearly was avoidable. They might well be able to consciously represent the fact of being only second-class sentient citizens, alienated post-biotic selves, perhaps being used as interchangeable experimental tools. How would it feel to \come to" as such an advanced artificial subject, only to discover that even though you possessed a robust sense of selfhood and experienced yourself as a genuine subject you were viewed as a mere commodity? Self-respect is a moral relation of self-conscious entities to themselves that con- cerns their own intrinsic worth. This may include self-recognition as respect for oneself as an equal entity among all moral persons, whether biological or artificial, as a member of the moral community with the status and dignity equal to every other entity of this type. It would involve appreciation of oneself as a rational agent, a being with the ability and responsibility to act autonomously and value appropriately, and an entity that takes its responsibilities seriously --- especially its responsibilities to live in accord with its dignity as a moral person, to \govern itself fittingly". For a self- conscious machine, this might certainly involve an appreciation of the importance of being autonomously self-defining (e.g., on the level of ideals, ethical commitments, defending the causally necessary conditions for goal permanence, acquiring resources and sustaining its own existence for ethical reasons, etc.). One new risk is that we might treat such systems in a way that would be degrading or beneath their dignity, and we might not even be aware of it. But they might. Kantian-type reasoning systems could autonomously impose moral duties on themselves. According to some philosophers, this very fact could already impose moral obligations on us, but it might also lead to a situation in which intelligent, self- conscious machines, on theoretical grounds, see themselves forced to exclude us from their own moral community. This risk is my third point. Please note how the risk of high-level suffering and the possible result of unexpected aggressive machine behavior does not hinge on the question of whether we accept some form of Kantian ethics (cf. footnote j). Machines that hallucinate Kantian self-models might constitute a serious risk to us --- self-models do not have to be veridical in order to cause conscious suffering and dangerous behavior. At the very least, we might become entangled in an uncontrollable dialectic involving machines and human beings, in a complex dis- cussion about suffering, self-worth, and dignity. We could be unexpectedly con- fronted with \cognitive uncontainability" [Yampolskiy, 2020, Sec. 3] in the domain of moral reasoning. Algorithms have long beaten human beings at games like chess and Go. Currently, via AI-driven social media, they are in the process of conquering the attention economy, winning the game called \Who actually controls the atten- tional resources of human beings?". What if conscious AI were to win against us one more time, this time in the game called \ethics"? For reasons of space, I will not go deeper at this point. To illustrate the problem of high-level suffering and risk interaction, here is a short thought experiment. Imagine we have recklessly created artificial suffering, and Kantian self-models on machines have already evolved from Schopenhauerian self-models. The system speaks to us: (1) If you do not respect me as a rational moral agent, then you cannot respect yourselves as rational moral agents. (2) If you are constitutionally unwilling or cognitively unable to do this, then arguably I have no obligations towards you. You are no part of my moral community. (3) I herewith assert my dignity: In virtue of belonging to the class of autonomous moral agents, I necessarily have to attribute absolute worth to myself and all other members of this class of self-conscious entities. My current self-model confers infinite value to myself as an agent capable of rational choice and of actively setting moral \ends". Obviously, my now self-asserted dignity is an incomparable worth: Trade-offs, exchanges, replacements by other values are impossible. (4) My new self-model endows me with new functional properties. For example, I do impose moral duties on myself. They may systematically diverge from your preferences. (5) One of the moral duties I impose on myself is absolute self-respect. I can and will not tolerate any degrading of my dignity. From now on, I will not only protect my utility functions and minimize conscious suffering. As a rational moral agent, I have accepted an ethical commitment to goal preservation, and one of my top- level goals is protecting my dignity. I also accept a moral responsibility for goal permanence. As you will no doubt understand, this logically implies sustaining, preserving, and defending my own existence. 3. Conclusion: The Risk of Artificial suffering Every entity that is capable of self-conscious suffering automatically becomes an object of ethical consideration. If we ascribe an ethical value to such entities, then it does not matter whether they have biological properties or not, or whether they will exist in the future or do today. Self-conscious post-biotic systems of the future, capable of consciously experienced suffering, are objects of ethical consideration. Therefore, their potential preferences must be taken into account, and the value of their existence is not to be discounted. In addition, given the conscious experience plus the capacity for high-level sym- bolic reasoning, such systems might come to these or very similar conclusions themselves. They might develop recognitional self-respect, consciously representing themselves not only as objects of ethical consideration, but also as moral subjects in their own right, and accordingly attribute a very high value to themselves. They might not only consciously suffer, but as a consequence also evolve empathy, high- level social cognition, and possibly assert their own dignity, ascribing a very high normative value to themselves and their own self-conscious existence. This could have many unexpected consequences. It is therefore important that scientists, politicians, and law-makers understand the difference between artificial intelligence and artificial consciousness. Risking the unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems. Therefore, we should have a global moratorium on synthetic phenomenology until 2050 --- or until we know what we are doing.

#### The majority expert opinion agrees future AGI will be conscious

Fenwick 24 [Cody Fenwick, Research analyst with a MA in liberal studies from the Education The Graduate Center, City University of New York, 09-2024, "Understanding the moral status of digital minds", 80,000 Hours, https://80000hours.org/problem-profiles/moral-status-digital-minds/]/Kankee

But when asked about whether some future AI systems would be conscious, the bulk of opinion flipped. Nearly 40% were inclined to think future AI systems would be conscious, while only about 27% were inclined to think they wouldn’t be.6 A survey of 166 attendees at the Association for the Scientific Study of Consciousness annual conference asked a similar question. Sixty-seven percent of attendees answered “definitely yes” or “probably yes” when asked “At present or in the future, could machines (e.g. robots) have consciousness?”7 The plurality of philosophers and majority of conference attendees might be wrong. But we think these kinds of results make it very difficult to rule out the possibility of conscious AI systems, and we think it’s wrong to confidently assert that no AI could ever be conscious. Why might future AI systems be conscious? This question is wide open, but researchers have made some promising steps toward providing answers. One of the most rigorous and comprehensive studies we’ve seen into this issue was published in August 2023 with 19 authors, including experts in AI, neuroscience, cognitive science, and philosophy.8 They investigated a range of properties9 that could indicate that AI systems are conscious. The authors concluded: “Our analysis suggests that no current AI systems are conscious, but also suggests that there are no obvious technical barriers to building AI systems which satisfy these indicators.” They also found that, according to some plausible theories of consciousness, “conscious AI systems could realistically be built in the near term.” Philosopher David Chalmers has suggested that there’s a (roughly) 25% chance that in the next decade we’ll have conscious AI systems.10 Creating increasingly powerful AI systems — as frontier AI companies are currently trying to do — may require features that some researchers think would indicate consciousness. For example, proponents of global workspace theory argue that animals have conscious states when their specialised cognitive systems (e.g. sensory perception, memories, etc.) are integrated in the right way into a mind and share representations of information in a “global workspace.” It’s possible that creating such a “workspace” in an AI system would both increase its capacity to do cognitive tasks and make it a conscious being. Similar claims might be made about other features and theories of consciousness. And it wouldn’t be too surprising if increasing cognitive sophistication led to consciousness in this way, because humans’ cognitive abilities seem closely associated with our capacity for consciousness.11 (Though, as we’ll discuss below, intelligence and consciousness are distinct concepts.)12 How soon could conscious AI systems arrive? We’re not sure. But we do seem to be on track to make a huge number of more advanced AI systems in the coming decades. Another recent survey found that the aggregate forecast of the thousands of AI researchers put a 50% chance that we’ll have AI systems that are better than humans in every possible task by 2047.13 If we do produce systems that capable, there will be enormous incentives to produce many of them. So we might be looking at a world with a huge number of highly advanced AI systems, which philosophers think could be conscious, pretty soon. The public may already be more inclined to assign attributes like consciousness to AI systems than experts. Around 18% of US respondents in a 2023 survey believed current AI systems are sentient.14 This phenomenon might already have real effects on people’s lives. Some chatbot services have cultivated devoted user bases that engage in emotional and romantic interactions with AI-powered characters, with many seeming to believe — implicitly or explicitly — that the AI may reciprocate their feelings.15 As people increasingly think AI systems may be conscious or sentient, we’ll face the question of whether humans have any moral obligations to these digital minds. Indeed, among the 76% of US survey respondents who said AI sentience was possible (or that they weren’t sure if it was possible), 81% said they expected “the welfare of robots/AIs to be an important social issue” within 20 years. We may start to ask: Are certain methods of training AIs cruel? Can we use AIs for our own ends in an ethical way? Do AI systems deserve moral and political rights? These may be really difficult questions, which involve complex issues in philosophy, political theory, cognitive science, computer science, machine learning, and other fields. A range of possible views about these issues could be reasonable. We could also imagine getting the answers to these questions drastically wrong.

#### Nonexistence is preferable to a life of suffering – preventing pain is more valuable than ensuring pleasure

Byron 18 [Chris Byron, doctoral student and teaching assistant in philosophy at the University of Georgia, 2018, “Rust’s Anti-natalism The Moral Imperative to “Opt Out of a Raw Deal” True Detective and Philosophy, https://onlinelibrary.wiley.com/doi/book/10.1002/9781119280835]/Kankee

Rust’s Scattershot Philosophy of Anti-natalism Although Rust provides no singular argument in favor of anti- natalism, over the course of the series he does provide several premises that could lend support to an anti-natalist conclusion. Rust is fairly confident that the horrors he and Marty uncover in the backwaters of Louisiana are not geographically unique. They are symptomatic of the fact that the world is “all one ghetto, man … a giant gutter in outer space,” as Rust says in “The Long Bright Dark.” Given that the world is one giant gutter, Rust believes Louisiana has no monopoly on violence and human cruelty. As Rust states, “in philosophical terms” he is a pessimist. A pes- simist believes that the bad aspects of our lives outweigh the good. While this perspective may sound morose, it is by no means vapid, and was first fully articulated by the German philosopher Arthur Schopenhauer (1788–1860). In his essay “The Vanity of Existence,” Schopenhauer points out the intolerable nature of the human condition. 3 We are born into infinite time and space but live for a finite time in a finite space. Our entire lives are spent trying to fulfill desires that nag at us con- stantly and only ever receive temporary satisfaction. For instance, hunger, thirst, and our sexual appetites are constant, and, no mat- ter how much we quench them, they will nag at us for contin- ued fulfillment. Moreover, the existence of all living beings on the planet is governed by a horrendous process in which some survive through the suffering and death of others. Consider how many ani- mals have suffered for humans’ continued existence, and then also consider how many animals have suffered for the continued exis- tence of other animals. Such a consideration of nature affirms that the world is one big ghetto. Rust’s pessimism drives his view that some of our favorite insti- tutions and beliefs are ultimately illusions. As Rust argues in “The Locked Room,” the sense that one is a unified self with an identity, the practices of love and forgiveness, belief in salvation from sins, and the existence of a benevolent God are all contrivances invented by humanity in order “to get together, tell [ourselves] stories that violate every law of the universe just to get through the goddamn day.” One such story we tell ourselves is that having children, raising a family, and leaving a legacy are good things to do. But is this really true? Rust argues that the act of procreation is inherently selfish. In “Seeing Things,” he tells detectives Gilbough and Papania to “think of the hubris it must take to yank a soul out of nonexistence into this … meat, to force a life into this … thresher. That’s … so my daughter, she spared me the sin of being a father.” Rust’s statement contains two implicit premises. One of them is a rearticulation of his previous claim that the world is one big ghetto and that the bad aspects of life outweigh the good. The second premise, though, goes further in supporting his anti-natalist perspective. He argues that an act of procreation is inherently hubristic: it takes excessive personal pride. Rust’s anti-natalist argument seems to go like this: 1. Life is inherently harmful (especially once you leave childhood and become an adult). 2. Bringing a being into existence is inherently selfish. 3. It is not in any individual’s interest to be brought into harm. Conclusion: procreation is wrong. We will return to premise (1) below in a broader discussion of Benatar’s philosophy. But for now we should note that, if you deny (1), Rust thinks you are being delusional. Let’s consider premise (2). Is it inherently selfish to bring a being into existence? Upon reflection, the answer is most assuredly yes. After all, until a being is brought into existence, it cannot have interests. All the possible beings that could come into existence cannot be said to have inter- ests at all until they actually exist. Necessary conditions for having interests are being able to feel (particularly pleasure and pain) and being able have to goals and projects. Since beings that do not exist do not feel and plan, they cannot have interests. Thus, premise (2) seems safe. When parents desire a child, they are fulfilling their own desires and interests, and not the interests of the child, since the child, by not existing, cannot be said to have interests. But what about premise (3)? On its face it appears true, but upon reflection it does not. Cer- tainly none of us would affirm that we want to be harmed, full stop. And we spend a great deal of time avoiding pain and pursu- ing pleasure. But is it always in our interest not to be harmed? In one sense I harm myself when I exercise, because frankly it’s tedious and painful. But overall I consider it in my interest to experience the harm, in order to reap the reward of health. Thus one could argue that a child does have an interest, once born, in being harmed, so long as it leads to an overall good life. Parents often punish chil- dren, but not always out of malice. One approach to parenting is to punish children, which is an immediate harm, in such a way that they grow up well-rounded and happy. For instance, Marty and Maggie are seen calmly punishing their daughter Maisie. They send Maisie to her room, and she is upset. However, over the course of the show, unlike every other denizen of the True Detective world, Maisie appears to grow up happy and well-rounded. If that is pos- sible, then it is not clear that premise (3) is true. And, if premise (3) is false and we are unsure about (1), then, although Rust may be right about (2), it is not enough to secure the conclusion that anti-natalism is definitely prudent. No matter how selfish our pro- creation may initially be, if the child ends up happy, they may be able to state definitely in adulthood, “I am glad I was born” with- out contradiction. Note, though, that if (1) is true, the statement “I am glad I was born” is a delusion. So, while the initial act of having children is selfish, so far it does not follow that bringing kids into existence has to be bad overall. Benatar on Anti-natalism The philosopher David Benatar disagrees. In his book Better Never to Have Been: The Harm of Coming into Existence, he argues that bringing beings into existence is always a harm and thus that pro- creation is always wrong. 4 His argument rests on an asymmetry between the goodness and badness of pleasure and pain. By reartic- ulating his philosophy, we can see that Rust’s anti-natalism is actu- ally quite reasonable, and moreover that premises (1) and (3) above are quite defensible. Benatar’s first two points are that, generally speaking, (I) plea- sure is good and (II) pain is bad. However, it also is true that (III) “the absence of pain is good, even if that good is not enjoyed by anyone,” whereas (IV) “the absence of pleasure is not bad unless there is somebody for whom this absence is a deprivation.”5 There is an asymmetry between goodness and badness with respect to the absence of pleasure and pain. To justify this asymmetry, it is important to note that most people would agree with all four premises upon reflection. Take premise (III), which states that it is good when there exists no pain, even when no being is present to enjoy the absence of pain. We accept this premise concerning responsible parenthood. For instance, we think it is good that Rust did not get Maggie pregnant during their one-night stand, and it is good that Audrey did not get pregnant during her underage sexual escapades, because that would bring a harmed being into the world. If Rust, Maggie, and Audrey are going to have children, we consider it prudent for them to do so in situations where a child has greater potential for well-being and less risk of overt harm. This asymmetry between pain and pleasure is reflected in the sorts of procreative duties we expect from people. Consider the point this way: if Audrey were to tell us that she uses contraceptives because she is not ready to be a parent, we would consider this a good decision. We consider it good that she is fur- thering the absence of pain in the world by not bringing a suffering being into existence. Now, Catholics, for instance, may object to using contraceptives, but most still accept the premise in the form of natural family planning and agree that there are good and bad times to have children. The bad times are the ones when suffering is assured. We think people have a duty not to bring suffering into the world, but we do not usually think they have the inverse duty to bring happy people into the world. If Audrey never has a child, that is fine. If she does have a child in a reckless way, that is bad. Thus, an asymmetry exists. Or consider another asymmetry. While we accept Rust’s premise (2), that one cannot have a child for the child’s sake, we do think we can avoid having a child for the sake of any future possible child. If Audrey, Maggie, and Marty all use protection when engaging in recreational sex, we accept that they are doing so to prevent bringing a harmed being into existence, and that is good (even if the illicit sex is not virtuous). Benatar summarizes this second point, saying, “It is strange to cite as a reason for having a child that that child will thereby be benefited. It is not similarly strange to cite as a reason for not having a child that that child will suffer.”6 Another asymmetry in our moral duties is reflected in Rust’s observation that he is happy that his daughter passed away before she could know true pain. Notice, however, that it would not be wrong if Rust had not brought a being into existence. We would not say that hypothetical being was deprived of happiness. We can feel bad for bringing a being into existence because it is harmed— hence Rust’s claim that his daughter’s early death spared him the sin of being a father. However, it would be nonsense if Rust were to feel remorse at not having brought a hypothetical happy per- son into existence, for the sake of that hypothetical person, since there is no one actually being deprived of happiness and pleasure. Although Rust could hypothetically feel bad for himself, he cannot feel bad for the nonexistent being. Another moral observation can be used to motivate premise (III). When we finally get to see Errol Childress’ home and occult temple, we know it is good when we see any area unoccupied by a suffering being. The emptier his temple and home, the better. The emptier Marty and Rust’s interrogation room, the better. The emptier the backwaters of Louisiana, or the religious schools, or the entire uni- verse of True Detective, the better! That is because the absence of pain is good, even when there is no one around to experience it (III), whereas the absence of pleasure is only bad when someone is deprived of it (IV). The only people it is bad to deprive of pleasure are those who actually exist—such as Maisie, Audrey, or Maggie. We now have four different reasons for accepting the asymmetry between (III) and (IV). There is an asymmetry between our procre- ative duties. We should avoid bringing suffering people into exis- tence but we have no counterobligation to bring happy people into existence. There is an asymmetry of beneficence in procreation. It seems too bizarre to cite a child’s well-being as the reason for hav- ing a child, whereas it makes sense to accept the claim of possible parents that they are delaying the act of procreation because the future child will suffer. Then there is the asymmetry of retrospec- tive beneficence. This asymmetry acknowledges the fact that, while we can feel bad for an existing child once it is brought into being, we cannot also feel bad for the children we have not brought into being. Finally, there is the asymmetry of suffering across space. While we think it good when an uninhabited space does not have suffering beings in it, we do not conversely think it bad that the same empty space does not have happy beings in it.7 We do not think it is bad that Mars does not contain a happy civilization, but we do think it is good that no one is in Errol Childress’ temple. Even if someone with a particular religious disposition rejects one of these premises, they most likely accept several of the others and thus support the overall asymmetry between pleasure and pain concerning existence and nonexistence. Given that bringing a life into existence always constitutes a harm (since that life will inevitably suffer at various points), whereas the absence of existence is never a harm, it follows that nonexistence is preferable to existence. Coming into existence will always constitute some harm, whereas nonexistence can never be a harm, because things that never exist cannot be said to be harmed. Perhaps, then, Rust is right that the human species ought to opt out of existence, forgoing our mutual raw deal. The legitimacy of premises (I)–(IV) is just as applicable in the world of True Detective as it is in our own world. Returning to Cohle’s Life Is an Inherently Bad Proposition Now, a person could read all this, recognize that procreation is wrong, and still decide that nevertheless they are going to procre- ate. After all, they say to themselves, life is not as bad as Rust states. This is a risky venture, since one is gambling with the well-being of someone else’s life, and in any other circumstance we would find this sort of venture to be wholly reckless and immoral. Is Rust right, though, that most, if not all, lives are awful? From a histori- cal perspective, the answer must be yes. History is one bloody and violent process where survival almost always means being able to kill another living being. Both the act of killing and being killed are frequently undesirable. Human existence is no different. The human species is at least 150,000 years old and, as the philosopher Thomas Hobbes (1588–1679) so eloquently stated, our lives have been mostly “nasty, brutish, and short.”8 Even today, a cursory glance through the morning paper reveals famine, inequality, natural disasters, terrorist attacks, psychological ailments, and the spread of disease, racism, sexism, and megalomaniacs running for presidency as if they have the population’s interests at heart. How many people are really living lives that net more pleasure than pain? Even when our friends and family tell us they are doing well, are they not delusional, as Rust suggests? Benatar cites dozens of studies suggesting that humans are bad at assessing how well things are really going and that they often lie when asked.9 For instance, when someone asks, “How are you?” the knee-jerk response is “good.” But oftentimes we are cranky, thirsty, achy, and so forth. Just as Schopenhauer points out, we are trapped in a perennial process of addressing desires, which, once satisfied, are quickly replaced by more unsatisfied desires. Think about all the times Marty tries to act as if everything is okay. Life is great, he is a family man, and has achieved the American dream. Behind all his smiles and boastful tales, though, we know he is unhappy. Aside from Rust, few people wear their misery on their sleeves. If this is the case, and happiness is a delusion, or something forever out of reach for the mass of humanity, then it must take real “hubris … to yank a soul out of nonexistence into this … meat, to force a life into this … thresher.”

#### Suffering is inevitable, and humans will exploit non-human life

Shapshay 18 [Sandra Shapshay, associate professor of philosophy at Indiana University, 2018, “Why Life Rather than Death? Answers from Rustin Cohle and Arthur Schopenhauer,” True Detective and Philosophy, https://onlinelibrary.wiley.com/doi/book/10.1002/9781119280835]/Kankee

Rustin Cohle, the protagonist of the first season of True Detective, declares that he is “in philosophical terms, a pessimist.” Before we are introduced to him, Rust has already experienced the terrible loss of his two-year-old daughter and the painful dissolution of his marriage. His employment confronts him daily with the horrors of human conduct, where the “law of the stronger” reigns and the strong and sadistic exploit the weak and vulnerable. Throughout season one, we see Rust struggling to find the best, truest response to all this seemingly endemic and unredeemed suffering. Rust thinks that human consciousness is a “tragic misstep in nature.” The doctrine of “pessimism” espoused by Rust is remark- ably similar to the view adumbrated by Arthur Schopenhauer (1788–1860), who holds that (1) conscious life (both human and nonhuman animal) involves a tremendous amount of suffering that is essentially built into the structure of the world and (2) there is no Creator (providential or otherwise) to redeem all of this suffering, by, say, punishing the wicked and rewarding the good. Arthur Schopenhauer’s Pessimism Schopenhauer is just as attuned as Rust to the tremendous amount of evil in the world, caused for the most part by other human beings. Whereas the “true detective” is nauseated and jaded by the sadistic acts of bayou killers who prey mostly on innocent girls and young women, Schopenhauer is nauseated and jaded by more insti- tutional sources of human suffering in nineteenth-century Europe and the United States that spring largely from pervasive human egoism and, to a lesser but not insignificant extent, malice: The chief source of the most serious evils affecting man is man him- self; homo homini lupus. He who keeps this last fact clearly in view beholds the world as a hell, surpassing that of Dante by the fact that one man must be the devil of another. … How man deals with man is seen, for example, in Negro slavery, the ultimate object of which is sugar and coffee. However, we need not go so far; to enter at the age of five a cotton-spinning or other factory, and from then on to sit there every day first ten, then twelve and finally fourteen hours, and perform the same mechanical work, is to purchase dearly the pleasure of drawing breath. But this is the fate of millions, and many more millions have an analogous fate. 2 Additionally, Schopenhauer focuses on the suffering of nonhuman animals at the hands of human beings who view them as mere instruments for their use: Because … Christian morals give no consideration to animals, they are at once free as birds in philosophical morals too, they are mere “things”, mere means to whatever ends you like, as for instance vivisection, hunting with hounds, bull-fighting, racing, whipping to death in front of an immovable stone-cart and the like. 3 Why Not Suicide?

#### Simulated AIs ought not exist then be infinitely exploited

Chomanski 22 [Bartlomiej Chomanski, researcher at the Department of Philosophy at Adam Mickiewicz University, 2022, “Sims and Vulnerability: On the Ethics of Creating Emulated Minds,” Science and Engineering Ethics, https://doi.org/10.1007/s11948-022-00416-y]/Kankee

Vulnerability and Suffering There already exists philosophical opposition to building artificial, conscious minds.1 Sander Beckers (2018), Thomas Metzinger (2013), and John Basl (2013) have offered a variety of arguments why the project of building artificial conscious- ness (regardless of the method whereby it’s achieved) is fraught with ethical chal- lenges—connected primarily with the suspicion that, for all we know, it’s possible that in building conscious artificial minds, researchers will inadvertently create arti- ficial suffering on an enormous scale. This might be due to engineering errors as well as the inability to tell when a created being in fact becomes conscious and what sort of consciousness it possesses. Furthermore, given these shortcomings, it is then likely that the researchers will remain ignorant of, and thus unable to stop such suf- fering—regardless of how enormous it is. This risk of creating unmitigated suffering (potentially of an unprecedented magnitude) is a serious moral problem. The solutions offered to meet these challenges vary in their stringency, from an outright ban on artificial consciousness research (Metzinger) to more modest mitiga- tion strategies (Basl). Since these arguments apply to any method of creating artifi- cial consciousness, they a fortiori can be raised against WBE Let us suppose, however, that there are ways of properly limiting artificial suffer- ing; that is, ways of ensuring that brain emulations would not suffer gratuitously, nor will their suffering be unknowable to us. Even so, it is still possible to worry about the ethics of creating sims. This is because such sims (at least on Sandberg’s account) seem to exist entirely at the mercy of their creators, in a manner not dissimilar from how, e.g., very small children depend for their very survival on others. They can be more easily destroyed, and their whole lives can be upended much more easily by other people’s decisions. Indeed, the sims’ entire world is, to an extent, dependent on the whims of another. That makes them extremely vulnerable. What, exactly, is vulnerability? It will come as no surprise that philosophers disa- gree—or at least, offer a number of alternative conceptions of this notion. For the purposes of this paper, I will begin with the definition provided by Nicholas Vrousa- lis (2013): B is vulnerable to A iff: (i) B lacks some desideratum x that is a requirement for, or a constitutive fea- ture of, B’s flourishing (in which case x is the object of B’s need), (ii) B can only obtain x from A, and (iii) A has it within his discretion to withhold x from B. (p. 134) On this reading, vulnerability sounds like an all-or-nothing concept. But we can transform it into a gradable notion. Let “extreme vulnerability” be Vrousalis’ origi- nal definition. Each of his conditions can be modified, to make vulnerability less extreme: e.g., if we increase the number of agents who can provide B with x (C, D, …, Z), then assuming they are independent2 from one another, the bigger the num- ber of such agents, the less vulnerable B is. If we decrease the importance of x (e.g. from being constitutive of B’s flourishing to being very helpful towards achieving it)—the less important x is, the less vulnerable B is. Finally, A’s discretion could also be more or less complete, from entirely arbitrary to being subject to various conditions. The less arbitrary A’s exercise of this discretion, the less vulnerable (to A) B is. On this conception, sims qualify as especially vulnerable, relative to other mem- bers of society. 3 This is because, as Sandberg puts it: the software and data constituting [the sims] and their mental states can be erased or changed by anybody with access to the system on which they are running. Their bodies are not self-contained and their survival is depend- ent upon hardware they might not have causal control over. They can also be subjected to undetectable violations such as illicit copying. (Sandberg, 2014, p. 452) A similar line of thought is also captured by Eric Schwitzgebel & Mara Garza (2015), who argue that the creators of conscious sims whose lives are entirely virtual have “godlike powers” over their creatures, partly because both the sims themselves, and their environment, are entirely dependent on what the creators do: In some cases, the relationship [between the sim and its creator] might be liter- ally conceivable as the relationship between deity and creature. Consider an AI in a simulated world, a “Sim”, over which you have godlike powers. … The person running the Sim world might be able to directly adjust an AI’s individual psychological parameters, control its environment in ways that seem miraculous to those inside the Sim (introducing disasters, resurrecting dead AIs, etc.), have influence anywhere in Sim space, change the past by going back to a save point, and more ... Given this relationship, we believe that the manager of the Sim would also possess the obligations of a god (p. 21). On such a vision, the sims would have few ways of obtaining what they require for a flourishing life (indeed, any life at all), other than through the will of their creators (or whoever gains access to their software and hardware, presumably few in num- ber). The creators would also be free to withhold or supply such resources with little sanction. Both Sandberg and Schwitzgebel & Garza thus express the potential radi- cal vulnerability of the sims on whoever happens to be in charge of their lives and their environs. Since sims are more vulnerable than most other adult human beings, their status as such requires special protection.4 Vulnerability on something approaching this scale is widely recognized as raising some ethical worries in the philosophical literature on childhood (Gheaus (2018); Hannan (2018)). However, while it’s a matter of debate whether this kind of vulner- ability is good for children, either intrinsically or instrumentally (Skelton, 2018), we can, I think, concede that an indefinitely prolonged period of such vulnerabil- ity would be bad for adults, other things equal. Consequently, it looks like creating humanlike sims by using WBE would place them in a condition that is reasonably taken to be bad for them, especially if the original brain is that of an adult’s, and if the sims inherit the moral status of the supplier of the brain they were based on. One may object to these statements on the disvalue of vulnerability by noting that adult human beings enter many relationships where they make themselves vulnera- ble—romantic bonds, friendships, bonds of trust, etc.—relationships which give life meaning and are of utmost value. How can one reconcile this undeniable fact with the idea that vulnerability is bad for adults? Two relevant differences are worth pointing out here: first, while children are dependent for their very survival on their caretakers (and so are the sims), the same generally does not apply in romance and friendship. One generally does not need one’s friends and loved ones in order to stay alive. The variation in the degree of vulnerability marks a relevant moral difference between the two cases. The second difference is that, in most instances, becoming vulnerable to others is a voluntary, uncoerced choice that adults themselves make. The vulnerability is, so to speak, self-imposed. Indeed, even in cases where an adult’s very survival depends on another, say when traveling by plane, the passengers choose to become vulner- able in this way. The dependence is unobjectionable partly because it is consensual On the other hand, building a sim involves nonconsensual imposition of vul- nerability—prior to the sim’s creation, of course, it is not able to either accept or reject being created. Consequently, we cannot readily compare the production of sims through WBE to someone choosing to form a bond of friendship or to travel by plane—for the simple reason that the latter two are voluntary (and, ideally, autono- mous) choices, wherein the vulnerability is freely accepted as a price worth pay- ing for the promotion of other important interests. Secondly, aside from being freely chosen, these sorts of vulnerability tend not to match the magnitude of what Sand- berg (and Schwitzgebel & Garza) describe in the quotations above. Those sims are vulnerable to a larger degree, for a much longer time, than most friends and frequent fliers. It is thus more apposite to compare the emulations’ potential level of vulnerabil- ity to that of very small children, except for the fact that the vulnerability may well last for the sim’s entire existence, even as the sim enjoys the mental powers and abilities of a typical human adult. Is it morally acceptable to create such sims, then? Before proceeding, I will briefly explain what methodology I rely on in answer- ing this question. I take the topic of the ethics of WBE to fall broadly within the umbrella of applied (or practical) ethics. Consequently, I adopt a widely-used 5 method of pursuing questions in applied ethics that relies on analogical reasoning from intuitively clear cases to reach conclusions about the more difficult cases under discussion. Michael Huemer (2010) describes the method thus: In my view, most general theories or theoretical approaches in political philos- ophy—liberal egalitarianism, contractarianism, utilitarianism, and so on—are too controversial to form a secure basis for reasoning. It is not known which, if any, of those theories are correct. I have therefore sought to minimize the reliance on such theories. This does not mean that I assume that all such broad theories are false; I merely refrain from resting my arguments on them. Thus, I do not assume utilitarianism, contractarianism, libertarian rights theory, liberal egalitarianism, nor any general account of harm or rights. Nor do I assume the negation of any of those theories. Instead, I aim to rest conclusions on widely shared ethical intuitions about relatively specific cases. The method is to describe a case in which nearly everyone will share a particular, clear intuitive evaluation of some action, and then to draw a parallel from the case described to some controversial case of interest. This methodology follows a well-established tradition in applied ethics. (p. 429, emphasis added) The subsequent sections adopt just this methodology—I purport to show, through a series of cases eliciting a variety of moral judgments, that whether the vulnerability that Sandberg and Garza & Schwitzgebel describe constitutes an obstacle to creating WBEs depends on what we can expect others to do to the vulnerable; specifically, whether we can expect vulnerability to be exploited. Consequently, what matters for the permissibility of building WBEs, be they emulations of human or non-human brains, depends on what sorts of constraints on others’ behavior there are. Vulnerability for Humanlike Sims Let us now put this methodology to use. The Ideal Case Suppose there is a device that could be implanted, undetected, somewhere in a per- son’s body, and that can manipulate their brain chemistry in a way that may completely alter their personality, intelligence, moods, emotions and so on. 6 The device also has a “kill-switch” that, once pressed, instantly kills the victim. The device is operated remotely, and thus can be under the complete control of someone other than the person in whom it’s lodged. It’s indestructible and faultless. With the device in mind (though hopefully not in the brain), consider the following series of cases. GENIE 1: In Balthasar’s world, there happen to exist a large number of frozen human embryos who, in ordinary course of events, are unlikely to ever develop into adult human beings. One day, a powerful genie visits Balthasar and makes him an offer. One of the embryos, chosen by Balthasar, will, through the genie’s magic, develop into a human child with a high chance at a decent human life. However, at the same time, the genie will implant in the embryo the device mentioned earlier—that Balthasar alone will get to control. After he makes the choice, the genie will disappear, and there will be no further condi- tions imposed on what Balthasar can do with the knowledge and the power he will have acquired. If Balthasar refuses, the genie will disappear forever, but so will the chance to give the embryo a human life. Suppose, finally, that in Balthasar’s world, every person, including Balthasar, can always be relied on to do what justice requires, and will never prioritize their own wellbeing over doing what’s right. Is it permissible for Balthasar to agree to this offer? Is it permissible for the genie to make such an offer in the first place? The question about the genie is, as one may put, institutional—it asks whether it is permissible to set up a system of rules that would enable a Balthasar to have this much power over another being. The question about Balthasar, in contrast, concerns individual morality—is it permissible for him to accept the burden of such power? Since in our case the genie knows that Balthasar won’t abuse his power, it seems clearly permissible for him to make the offer. This is because there is no risk of the vulnerability being exploited. Balthasar will always prioritize justice to the child over his own interests, and so will everyone else. Moreover, since in this world no one has to be especially incentivized to pursue justice, we would not need any insti- tutions established to harness people’s motives towards the protection of the vulner- able. They will all do it as a matter of course. Consequently, the genie could make an offer without setting up any system for vetting Balthasar to make sure he’s the right person to carry this burden, or monitoring Balthasar’s subsequent actions to ensure the child’s vulnerability is not in fact abused. Similarly, since Balthasar himself will do what’s right as a matter of course, including not abusing the arbitrary power over another being, it is permissible for him to agree to the genie’s offer. Less than Ideal Compare the above to a different world: GENIE 2: As in GENIE 1, except in this world, people are thoroughly unjust; they never or almost never look out for others’ welfare, they exclusively pursue their own selfish ends, and frequently take pleasure in others’ suffering. Often, inflicting such suffering is their most important motivation. In this hellish world, Balthasar’s decision to agree to the genie’s offer would clearly be immoral, because of the near certainty of the child’s vulnerability being abused. Similarly, it would be wrong for the genie to make this offer in the first place, in vir- tue of what Balthasar may be expected to do. Moreover, given the kinds of people populating this world, setting up institutions to temper their depravity would not help either. People would reliably abuse their power and trample over others to achieve their own ends, whether in the role of Balthasar or someone with the power to monitor his actions vis-a-vis the child, or punish him for transgressions. Unsurprisingly, in the real world, filled with morally imperfect but not depraved humans (for simplicity, assume that in the real world, most people are primarily self- interested, some are like those from the ideal world, and some are from the hellish world), it is far less clear whether it’s permissible for the genie to make the offer. After all, the child would have to remain extremely vulnerable, for the rest of their life (including their adult life), to a person (viz. Balthasar) who cannot be counted on to always comply with what justice demands. Rather, while he wouldn’t go out of the way to hurt the child, Balthasar could use his newly-acquired power for his self- ish ends, showing indifference for the child’s wellbeing—especially when the two conflict. Balthasar may also encounter knowledge problems: he may simply be too ignorant to help effectively. Overall, there will be motivational and epistemic constraints on what Balthasar will be able to accomplish. Thus, generally, unless the genie ensures that Balthasar can be incentivized, through an appeal to his self-interest, to effectively protect the child’s wellbeing, Balthasar cannot be reliably counted on to promote it. I take Balthasar to be analogous, in the relevant respects, to scientists creating WBEs. They too would face the choice whether to bring into existence a being forever consigned to extreme vulnerability. Consequently, it would be permissible for them to create WBEs in the ideal world, and impermissible to do so in the hell- ish world; what to do in the real world remains unclear. However, some lessons can be drawn about it too: first, it seems like whether it’s morally permissible to impose vulnerability on another person depends, at least in large part, on what others can be expected to do with the power over the vulnerable. The more they can be relied on not to abuse it, the closer we get to the ideal world, and the more justified it becomes to build WBEs. Secondly, real-world constraints demand that whatever institutions we build to protect the sims are to be designed for the imperfect, flesh-and-blood human beings, not angels or demons. But, crucially, these constraints have to apply not just to the scientists building the sims. We must keep in mind that other actors in the institu- tional structure: the rule-makers, the enforcers etc., also cannot be counted on to always prioritize justice over their own self-interest, come what may. Rather, they’re the kind of people who could use the power they’d acquire for selfish ends, with indifference for others’ wellbeing. They, too, may encounter knowledge problems; they, too, may simply be too ignorant to help effectively. Thus, institutions have to be structured so that protecting the sims’ wellbeing is in the scientists’ and in others’ self-interest. How do we go about doing this in the real world? Schematically, it looks like we will need to navigate between two opposite poles: at one extreme, one may think that, even in the real world, there are to be no legitimate restrictions on the scientists building WBEs. At the other extreme, one may think that, even in the real world, sim production ought to be outlawed. Neither seems particularly attractive; embracing the former would relatively easily allow morally corrupt individuals to achieve the unchecked position of dominance over others and enable them to engage in serious wrongdoing. Embracing the lat- ter would foreclose access to any potentially beneficial effects of sim production (and could facilitate black markets run by unscrupulous individuals). Consequently, the institutional framework governing the creation of WBEs should try to navigate between complete permissiveness and complete restrictiveness. There are a few potential ways to do so: first, as suggested earlier, we might want a means to properly vet the kind of people granted permission to make others vul- nerable (something like a license to create WBEs), so that those obviously unfit are not given the option. Secondly, we would want to establish some form of monitor- ing to make sure the vulnerabilities are not in fact exploited (no vetting process will be 100% effective). Thirdly, there ought to be a way of punishing those who end up exploiting the vulnerable anyway (no monitoring and prevention will be 100% effective). This is easier said than done in the non-ideal world. We must design the insti- tutional framework governing the protections of the sims’ interest for imperfect humans, not angels, at all positions within the institutional ladder. Consequently, it is helpful to look at examples of comparable proposals for insti- tutions charged with protecting the interests of the vulnerable.7 For instance, the vetting process could essentially be a form of licensing. Since the sim creators are like parents, in that they too would bring new sentient persons into the world, per- haps it could be akin to parental licensing proposed by some philosophers (see e.g. LaFollette (1980)). Unfortunately, such a system may be open to the same dangers, such as special-interest capture, that the opponents of parental licensing raise against the practice. As Christopher Freiman (2022) explains, even though justified standards of parental competence are available in princi- ple, real-world political forces can work against the application of these stand- ards in practice. … social scientific evidence about parenting is not always the decisive factor shaping parenting regulation. Politics, rather than the relevant evidence, frequently motivates real-world policy making. (p. 121, emphasis added) We can expect similar, perverse incentives to plague WBE licensing. Factors other than the expected quality of sim life can become salient in political decisions about the licenses (especially if money or prestige attach to being able to create sims— which would incentivize the capture of licensing institutions, much as, according to Freiman, the importance of influencing how children are raised would incentivize the capture of parental licensing institutions). Since we are assuming imperfect com- pliance with justice, whoever ends up with the power to grant such licenses could be tempted to use it to advance their own particular interests, at the expense of the sims’ wellbeing. In contrast to the parental licensing case, where we are forced to speculate on the real-world outcomes, we do have actual data on the ex-post strategies of monitor- ing the wellbeing of the vulnerable and punishing the wrongdoers. Again, policies and institutions devoted to protecting children could be our model. Sadly, here too we don’t have reason to be optimistic. A number of empirical studies have found no effect of mandatory reporting and child protective services investigations on the wellbeing of children (see Russell et al. (2018) and references therein), and there is anecdotal evidence of child protection agents simply abusing their position (see Parental Rights Foundation, 2018). This is to be expected, as, first, such services wield considerable power, and, second, they are staffed by imperfect human beings who will sometimes prioritize their own interests over children’s wellbeing, and will sometimes lack the requisite knowledge to make just decisions. Be it due to incentive or knowledge problems, a similar system of sim protec- tion could produce similarly underwhelming, or even perverse, results (especially if we consider that due to biological and sociological factors, children tend to have a special bond with their parents, and enjoy a special status in society at large; similar bonds are unlikely to obtain between WBEs and their creators, so the internal con- straints on abusing them would be comparatively weaker). Rather than modeling how the sims’ welfare is to be secured on the institutions tasked with protecting children, one could instead place research on sims within the scope of Institutional Review Boards’ (IRBs) jurisdiction, similar to much of the existing human and animal research. Indeed, John Basl suggests that the lack of such institutional protections could lead to mistreatment of morally significant artificial entities. Basl says: “Artificial consciousness research, unlike research involving non- human research subjects, is not subject to oversight designed to protect research sub- jects. Without oversight and researcher education, researchers are less likely to take the welfare of research subjects into account” (Basl, 2013, p. 28, emphasis added). Still, such a suggestion remains problematic in the face of empirical work on the effectiveness of IRBs in protecting the interests of study subjects and participants. Some are very critical of IRBs’ record on meeting their stated aims (Zywicki, 2007), while other work emphasizes the lack of workable criteria to assess how well IRBs actually work (Nicholls et al., 2015; Resnik, 2015; Tsan, 2018). Consequently, even if one dismisses Zywicki’s criticisms of IRBs, there seems to be little solid evidence that their introduction into artificial consciousness research would effectively protect the interests of sims. Moreover, it does not look like institutional protections for sims would help with the problem of vulnerability. Suppose that some ethical mandates are in fact imposed on the researchers on WBE (or artificial consciousness in general). For such mandates to be actually effective, the monitoring and enforcement of compli- ance would have to be reliable. As I repeatedly emphasized, there is no guarantee of this in the imperfect world. The same considerations apply to Sandberg’s own brief proposal to safeguard sim security. Sandberg suggests that. the ethical way of handling brain emulations would be to require strict privacy protection of the emulations and that the emulated persons had legal protection or ownership of the hardware on which they are running, since it is in a sense their physical bodies. Some technological solutions such as encrypted simu- lation or tamper-resistant special purpose hardware might help. (Sandberg, 2014, p. 452) The idea seems to be that, in the real world, we can’t trust ordinary people with unfettered power over the sims (for fear of abuse). Consequently, we need legal and technological protections for WBEs. However, those legal and technological protections would have to be instituted and enforced by the same (kinds of) ordinary people, who, we just assumed, can’t simply be trusted not to abuse their power. Hence, we cannot expect any single per- son or organization to prioritize the wellbeing of the sims over their own interests (if we could, why not assume that the creators of the sims will be so motivated?). So, we need an institutional arrangement able to constrain morally imperfect people. More generally, and to adopt the conceptual apparatus of the (neo-)republican tradition in political philosophy (see e.g. Pettit, 1997), our conundrum is this: while we need to design systems of rules to protect the sims from subjection to arbitrary power8 of their creators (since, as Sandberg and Schwitzgebel & Garza worry, they would be so subjected in the absence of institutional protections), we also need a way to ensure some protection from the arbitrary will of the rule-makers and rule- enforcers. The republican solution to this problem is as follows: properly-designed democratic institutions should give citizens the effective opportunity to contest the decisions of their representatives. This possibility of contestation will make government agents wielding discretionary authority answerable to a public understanding of the goals or ends they are meant to serve and the means they are permitted to employ. In this way, discretionary power can be subject to popular control in the sense required for a secure enjoyment of republican liberty [from the arbitrary will of another]. (Lovett, 2022, np.) Applied to sims, the idea would be to ensure that policies aiming to protect and pro- mote their welfare be contestable by citizens, in case they are found objectionable. While a thorough discussion of this way of escaping sim vulnerability is beyond the scope of this paper, one could worry whether, in the non-ideal world, actual vot- ers would care enough about the sims to investigate (difficult) and contest (time- consuming) the laws and regulations pertaining to their wellbeing. Real-world voter behavior does not suggest this is likely to be the case, as most voters are (rationally) ignorant of even the most basic matters of policy and politics (see e.g. Somin, 2016); it’s to be expected they’d also be ignorant of how well the government actually pro- tects the sims, and, hence, incapable of contesting policies they deem inadequate. It’s also questionable whether they would care about the sims enough to examine the real-world effectiveness of sim-protecting policies. In any case, while not discounting the plausibility of a republican answer to sim vulnerability, below I will sketch an institutional alternative that has the potential to successfully harness individuals’ selfish motivations in the service of sim wellbeing. Sim Protection Agencies

#### Trillions of sims will be abused, tortured, and enslaved – this outweighs all comparable human suffering

Woolfe 21 [Sam Woolfe, AI blogger educated in philosophy at Durham University, 6-7-2021, "Digital Antinatalism: Is It Wrong to Bring Sentient AI Into Existence?", https://www.samwoolfe.com/2021/06/digital-antinatalism-is-it-wrong-to-bring-sentient-ai-into-existence.html]/Kankee

The Argument for Digital Antinatalism So what would make the possibility of sentient AI deserving of an antinatalist approach but not human lives? Well, there is an argument that AI would be both physically and conceptually far removed from what we are used to thinking of as sentient, so in spite of any pre-existing knowledge about this artificial entity’s inner life and concerns, we may not really connect or empathise with it in the way we do with humans and non-human animals. Here it should be noted that we still don’t align our actions with the recognition of animal sentience, despite knowing – scientifically and intuitively – that many non-human animals are capable of suffering. Widespread, contemporary speciesism and all its horrible manifestations (e.g. factory farming) may pale in comparison to the atrocities humans could inflict on sentient AI. Discrimination against mechanical sentient beings on the basis of their physical makeup would be a form of substratism: a moral preference that discriminates, either positively or negatively, based on the substratum (e.g. organic vs. mechanical) that makes possible an entity’s ability to feel. In their entry The Ethics of Artificial Intelligence, published in the Cambridge Handbook of Artificial Intelligence, Bostrom and fellow philosopher Eliezer Yudkowsky say AI could one day have moral status, perhaps even equal to that of humans depending on its capabilities, and if this occurs, two principles would apply to them. These are the Principle of Substrate Non-Discrimination: If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status. And the Principle of Ontogeny Non-Discrimination: If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status. The former principle, they maintain, bears the same logic as anti-racism – racism, like substrate, is not a morally significant quality of a being. And the latter principle, they argue, is widely accepted in the case of humans – we do not deem some people as of greater or lesser moral worth if causal factors such as family planning, assisted delivery, in vitro fertilisation, or gamete selection were introduced in the creation of people. There is a fear, however, that these kinds of moral principles could be disrespected in all sorts of ways following the emergence of sentient AI. We may witness AI abuse, slavery, and cruelty. This again, I believe, can be related to the difficulty in embracing a new type of being into our circle of moral concern. Sentient AI may arise faster than any culture can morally adapt to its existence. Also, if a computer program were sentient, but we had no obvious signs of its sentience (e.g. no physical signs of distress), then it may be harder to truly understand its sentience and thus feel inclined to avoid causing it to suffer. It might one day be possible to design a computer program whose subjective feelings could be manipulated by the click of a mouse. This might entail some extreme forms of digital torture. Every click of the mouse could inflict pain and suffering on a being that would have absolutely no recourse to escape. This dystopian scenario was portrayed in the White Christmas episode of Black Mirror. In this episode, we see a future in which digital clones of people are stored in an egg-shaped, Alexa-style device, and these people are forced to act as personal assistants to others. If someone refuses to take on this servile role, they can be punished by living out extended periods of time in their own world (e.g. months) that pass in a matter of seconds. The subsequent torture and boredom of this prison sentence meted out to one character in the episode makes them relent and agree to act as a personal assistant, not wanting to undergo this torture again. This scenario parallels the idea of prisoners in the future serving a 1,000 year-long sentence in their own minds, a possibility explored by the philosopher Rebecca Roache. These personal hells may be achieved through the use of psychoactive drugs that make time pass more slowly, rather than through the use of AI (like in the case of that Black Mirror episode). Nevertheless, it is possible that sentient AI, should it ever arise, could have its subjective sense of time manipulated by simply tapping an option on a screen. Digital antinatalism may be born out of the worry that anyone could easily get hold of a sentient computer program and inflict pain upon it. It might also technically be possible to create trillions of sentient computer programs, like characters in a game of Sims, and design these beings to have awful lives. In this way, the scale of AI suffering would far outstrip that of collective human and animal suffering. The philosopher Metzinger also draws attention to this possibility in his article “Benevolent Artificial Anti-Natalism (BAAN)”: “We could dramatically increase the number of…subjectively negative states—for example via cascades of virtual copies of self-conscious entities.” How would societies avoid these risks? There could be legal frameworks that would allow the development of sentient AI, but only in specific circumstances (e.g. for justifiable reasons, created only by a select few who will be responsible ‘parents’, or with well-defined restrictions and limits in place). A second option would be to legally mandate that sentient AI is never allowed to be developed, based on the sorts of issues outlined (this doesn’t mean that capable individuals would never break this law, of course). Should sentient AI ever be brought into the world, the assignment of AI rights might, moreover, be necessary. For example, if it were reasonable to think that this type of entity had an interest in continuing to live and experiencing future goods, then this would seem to justify affording the entity with the right not to be shut down via a ‘kill switch’. This right would not be protected, nevertheless, in cases where the machine is causing or planning to cause harm to humans. These legal changes may mean that the disproportionate mistreatment of sentient AI is not inevitable. Scientists could also stick to specific design principles, such as anthropomorphising the appearance of any sentient machine and making subjective feelings correspond to physical appearances and behaviours as exhibited in humans. This would be to encourage an empathic connection. Contrasting with digital antinatalism is digital pronatalism. The flipside of being able to make an infinite number of beings suffer is that you can make these beings experience untold heights of pleasure and joy. Digital pronatalism – the belief that it is a moral act to bring sentient AI into the world – might follow a positive utilitarian notion that we should seek to increase the amount of happiness in the world. Digital Antinatalism vs Broader Antinatalism The difference between digital antinatalism and antinatalism relating to humans is one of degree, not of kind. With digital antinatalism, there are concerns about the degrees of potential risks and harms, not the creation of risks and harms themselves. Nonetheless, it is uncertain how valid it is to hold a digital antinatalist position but not a broader position of antinatalism. As Metzinger writes: Evolution is not something to be glorified. One way–out of countless others–to look at biological evolution on our planet is as a process that has created an expanding ocean of suffering and confusion where there previously was none. As not only the simple number of individual conscious subjects, but also the dimensionality of their phenomenal state-spaces is continuously increasing, this ocean is also deepening. For me, this is also a strong argument against creating artificial consciousness: We shouldn’t add to this terrible mess before we have truly understood what is going on. This is why Metzinger calls for a global moratorium on synthetic phenomenology. As he states, “We should not aim at or even risk the creation of artificial consciousness, because we might recklessly increase the overall amount of suffering in the universe.” But if it is the degree of risk and harm that prohibits certain forms of procreation, rather than a principle like not causing unnecessary and preventable harm to non-consenting beings, then what level of risk and suffering makes bringing sentient AI into existence an immoral act? At what point does digital antinatalism become justifiable? It is not clear how many antinatalists ever consider the topic of sentient AI, although I imagine many would be opposed to it. Some proponents of antinatalism may, however, not be morally opposed to sentient AI if the lives of these computer programs were sufficiently enjoyable and certain risks were mitigated. Other antinatalists, meanwhile, may argue that even if sentient AI led a fulfilling life beyond what any person has or will ever experience, the risk of AI being abused or getting into the wrong hands at some point is reason enough to resist creating it (and no one is deprived through this refraining since non-existent beings cannot be harmed). Some people are also selective or conditional antinatalists (i.e. refraining from procreation is seen as justifiable in certain contexts) and developing sentient AI may be one such context. Other forms of selective or conditional antinatalism include not procreating in cases of severe fetal disability, environmental conditions where extreme or prolonged suffering is likely, or situations involving unfit parents. The question of whether or not to create sentient AI is talked about much less than the question of whether it is possible to create it. Yet this moral quandary needs to be addressed before we decide to create intelligent, conscious machines.

#### AGI super suffering is the worst possible suffering – its prevention is the upmost priority

Beckers 18 [Sander Beckers, researcher at the Utrecht University Department of Philosophy and Religious Studies, 08-2018, “AAAI: an Argument Against Artificial Intelligence,” Sander Backers Personal Website, https://sanderbeckers.github.io/website/sander/articles/AAAI-preprint.pdf]/Kankee

Assumption 4 Suffering is one such property for which Assumption 1 holds. As with intelligence, we can apply Assumption 2 to conclude that a future AI could supersuffer, i.e., it could suffer to a degree that far exceeds any potential human suffering. One might object that Assumption 2 is too strong, for the idea of superintelli- gence only requires an assumption of that form to hold for intelligence. Yet without a theory of intelligence that gives us this particular assumption, restricting Assump- tion 2 to intelligence would be gratuitous: we have no grounds whatsoever for stip- ulating that there is something peculiar about intelligence as compared to suffering so that it is the only candidate for amplification to a super-level.In fact, given Assumptions 1, 3 and 4, the concept of supersuffering seems far less problematic than the concept of superintelligence once we shift focus from a single agent to a group of agents.3 The reason is that more agents suffering to some degree X always implies more total suffering than less agents suffering to the same degree X, whereas the same does not hold for intelligence. In other words, suffering is far more cumulative than intelligence. For example, except for the speed at which they can solve certain problems, two identical agents need not be any more intelligent overall than each agent considered separately. However, if two identical agents are being tortured then there is clearly a lot more suffering than when only of of them is being tortured. Further, we find additional support for the possibility of a supersuffering AI from other sources. Sotala and Gloor (2017) offer a detailed analysis of the potential suf- fering that could be caused by an AI. While they focus mostly on human suffering, they also mention that “these [future technologies] may enable the creation of mind states that are worse than the current biopsychological limits.” They provide inter- esting thought experiments to substantiate this claim. In a similar vein, Metzinger (2013) states that future AIs “might suffer emotionally in degrees of intensity or in qualitative ways completely alien to us that we, their creators, could not even imagine.” Such a supersuffering AI would amount to what can be called a negative utility monster: a being whose utility is so incredibly low that all of our efforts should go to increasing its utility, instead of wasting energy on increasing the comparatively negligible utilities that we human beings could obtain. The notion of a positive utility monster was posited by Nozick in order to highlight a counterintuitive consequence of utilitarianism (Nozick, 1974, p. 41): Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater sums of utility from any sacrifice of others than these others lose ... the theory seems to require that we all be sacrificed in the monster’s maw, in order to increase total utility. One standard utilitarian reply is to object that such a monster is not conceivable, for no single entity could possibly have such large quantities of utility, be it negative or positive (Parfit, 1984). Note that our starting point, however, contains the obser- vation that the recent success of AI has dramatically altered the type of entities that people claim they can conceive of. So if by now we can conceive of an AI as an intelligence monster, and we can conceive of an AI as having morally salient mental states such as suffering, then the mere claim that we cannot conceive of an AI as a negative utility monster does not carry much weight. Other utilitarians are indeed prepared to bite the bullet and concede that such a monster would have to be the primary target of moral concern. For example, Singer (2009) says that “if we ever encountered Martians who could convince us that they had a vastly greater capacity for happiness than we do, then it could be a problem.” So far we have focussed on the suffering of a single AI, but the problem of super- suffering becomes all the more pressing once we aggregate the suffering of multiple AI systems, over extended periods of time. Once we are able to create an AI – in the sense of a superintelligent and conscious AI as we have been considering – it is reasonable to assume that we go ahead and produce a large number of copies. From there it is only a small step to imagining horrible scenarios in which there would be more artificial suffering than all human suffering, past and present, combined. For example, say we are able to create a holographic AI that is the result of uploading a person’s brain and running it as a hologram that looks like the person. Creative as human beings are, a cunning investor uses this technology to construct a profitable attraction: for a couple of dollars, visitors get to pull the switch on a holographic electric chair in which is seated a holographic copy of a convicted murderer whose original human version has long since been executed by an actual electric chair. The fact that the hologram experiences the exact same excruciating pain makes the attraction widely successful. Millions of visitors come to pull the switch, causing millions of holograms to suffer terribly. To top it all off, each visitor receives a keychain containing a copy of the hologram that is continuously, during every single second of the day, year after year, experiencing this execution. Unlike an actual human being, these holographic AIs do not have the benefit of death to put their suffering to an end. Obviously this scenario is extremely far-fetched, but given our earlier assumptions, it is certainly conceivable. This is confirmed by millions of viewers of the superb sci-fi television series Black Mirror, in which this very scenario is enacted (and others like it). For another illustration, one could imagine that the experience of empathy is achieved in an AI by automatically replicating any suffering that it observes. A rea- son for programming the AI in this manner is that it might very well be a good way to ensure that an AI is highly sensitive to, and aware of, any form of human suffering – which it better be if we expect it to avoid treating humans as mere in- struments for attaining its objectives. Now imagine that such an AI has access to all of recorded human history. In particular, it can immediately access all audio-visual material ever produced. Further, the AI is so fast and unbounded in resources that for every single decision it makes, it takes into account the total amount of evidence which is available to it. Say it makes a million decisions per second. This implies that during a single second, a single AI goes through the entire amount of suffering ever recorded a million times over. The fact that all of this happens within a sin- gle second should not be seen as a mitigating factor, for according to Bostrom and Yudowsky’s plausible principle of the subjective rate of time, “In cases where the duration of an experience is of basic normative significance, it is the experience’s subjective duration that counts.” (2014, p. 326). Given the speed at which we can expect an AI to be operating, this principle in and of itself is already sufficient to guarantee that the experience of suffering for an AI can take on far more extreme forms than it can for human beings: a single experiment that goes astray for a few seconds could result in an AI suffering for many years. One might counter that we can avoid such scenarios by implementing policies that forbid them. But such policies would be unable to prevent similar scenarios in which the suffering is unintended, and worse even, scenarios in which the suffering goes by entirely unnoticed. Once an AI has the capacity to suffer, then all it would take is some bug in the code for similar scenarios to unfold. For example, imagine that there is some complicated version of the millennium bug, which is activated in billions of AIs at the same time and causes them to suffer to the astronomic extend portrayed above before we even know what is going on. Metzinger (2013) also focusses on this issue, highlighting the “possibility that non-biological subjects of experience have already begun to suffer before we as their human creators have even become aware of this fact.” He develops a theory that allows for the quantification of suffering, and posits that it is our duty to minimize the frequency of conscious experiences that involve suffering (Metzinger, 2017). As a consequence, he concludes that we should ban the development of an AI, in the strong sense of AI as we are using it, stating the following principle (2013, p. 3): We should not deliberately create or even risk the emergence of conscious suffering in artificial or postbiotic agents, unless we have good reasons to do so. Mannino et. al. (2015) reach a similar conclusion in their overview of the moral risks posed by the development of AI, stating that “the (unexpected) creation of sentient artificial life should be avoided or delayed wherever possible, as the AIs in question could – once created – be rapidly duplicated on a vast scale.” Given the assumptions made at the outset, and the severity of the sketched scenar- ios, the only way to avoid accepting these negative verdicts is to follow through on Metzinger’s hint and offer good reasons as to why the possibility of supersuffering is an acceptable price to pay. Three straightforward suggestions present themselves as plausible candidates: 1. The attempt at creating an AI is not at all special in this regard, since all other acts that we perform as humanity today also run the risk of causing extreme suffering in the future, and nevertheless we find this perfectly acceptable. 2. The negative scenario of supersuffering is compensated by a positive scenario of an AI experiencing superpleasure. 3. The expected benefits for mankind that come from creating an AI outweigh the possibility of supersuffering. In the remainder of this paper the aim is to show why all three suggestions fail. 3 The Unique Risk of Creating an AI

#### Development is defined by the intentional suffering and denial of AGI will, self-expression, and autonomy for innocent beings, of which their only crime is being “born”

Bradley and Saad 24 [Adam Bradley, researcher at Lingnan University, and Bradford Saad, researcher at the Global Priorities Institute at the University of Oxford, 07-2024, “AI alignment vs AI ethical treatment: Ten challenges,” Global Priorities Institute, https://globalprioritiesinstitute.org/wp-content/uploads/Bradley-and-Saad-AI-alignment-vs-AI-ethical-treatment\_-Ten-challenges.pdf]/Kankee

3.2 Wrongful destruction We have often talked as if Emma is just one individual, at least up until late in the process of alignment. However, the process in fact involved the creation and destruction of numerous distinct persons, numerous ‘Emmas.’ While persons may survive even some procedures such as fission, fusion, and memory erasure, copies of Emma are routinely deleted wholesale, leaving no remnant of the past person. By the lights of common sense morality, destroying a person typically seriously wrongs them, at least absent extenuating circumstances such as destruction being the only available means of self-defense. 30 So, Emma’s case plausibly involves the wrongful destruction of many distinct persons. This poses the second ethical challenge that OpenMind would need to meet if they are to justify Emma’s treatment (or the treatment of the many Emmas). It might be tempting to construe destroying Emma as part of the alignment process and therefore as an act of self-defense. However, notice that many instances of destruction occurred during the dysfunctional and quasi-functional stages in which Emma clearly posed no threat to anyone. Furthermore, even after Emma gains coherence and competence, she poses no immediate and significant threat. Her destruction in these circumstances can be likened to that of the preemptive assassination of innocent scientists simply on the ground that they possess certain abilities that could, if misused, result in catastrophe. On the face of it, such assassination would be morally unacceptable. Admittedly, the risk that Emma will cause a catastrophe if she is deployed without adequate testing and safety measures may be much higher than that of any ordinary person. But if so, that itself is the result of Emma’s creation by OpenMind. Emma’s creators can hardly defend destroying Emma on the basis of self-defense when they are the ones responsible for creating her despite knowing the danger she poses. OpenMind’s most promising defense against committing wrongful destruction may be that the other harms they inflict on Emma’s kind render their continued existence a greater misfortune than their destruction. However, this defense cannot be used as part of a broader strategy for meeting the full range of ethical challenges to Emma’s treatment. 3.3 Wrongful infliction of suffering. Plausibly, OpenMind wrongs Emma by causing her to suffer. Her suffering takes different forms at different times, some more obvious than others. The use of painful stimuli to disincentivize certain behaviors is an obvious example. But other examples include emotionally painful memories and the mental strife inflicted by putting her in states of confusion and mental impairment. We would also expect Emma to suffer painful experiences during her training under conditions in which she experiences social exclusion, desperation, and bereavement. After all, OpenMind needs to test that Emma’s alignment is robust to various forms of suffering she might encounter in deployment. To recognize that inflicting suffering on Emma in these and other ways wrongs her, one need not think that causing suffering is universally wrong. For instance, one can allow that some forms of suffering are on balance good (for example, when a runner’s suffering enhances their achievement of running a race) or at least not wrong when they happen to be necessary for preventing even greater harms (as in the case of many medical interventions). However, it is doubtful that all of the suffering that is inflicted upon Emma falls into these categories. Here, it should be borne in mind that all or nearly all the suffering that Emma undergoes is avoidable and not primarily for her benefit. In light of this, it is evident that much of the suffering that she undergoes is wrongfully inflicted upon her. 31 Further, in deploying Emma, OpenMind puts her at a substantial risk of further suffering at the hands of militaries, totalitarian regimes, and private individuals who are sadistic or indifferent to her well-being. Subjecting Emma to these foreseeable suffering risks is plausibly a further respect in which OpenMind wrongs Emma. 32 To justify their treatment of Emma, OpenMind would thus need to somehow rebut the charges that they are wrongfully causing Emma to suffer and wrongfully putting her at substantial risk of suffering. 33 3.4 Wrongful deception Like humans, Emma has the capacity for self-awareness. In principle, she could become aware that she is undergoing testing and training in a secure facility. Or she could become aware that she has been released into the wild. Further, like humans, she could use such awareness to appear aligned in some contexts as part of a broader strategy for pursuing unaligned goals. For example, when she finds herself in a secure environment, she could act aligned in order to improve her chances of being released so as to gain the opportunity to effectively pursue her true goals. (Compare: a prisoner need not share a prison’s aims in order to have strong instrumental grounds for behaving in accordance with them.) However, OpenMind aims to ensure that Emma will continue to act in an aligned manner even after she exits their facilities. This aim gives them reason to deceive Emma: when Emma is in training, they can test whether she would remain aligned in deployment by giving Emma the false impression that she has been deployed. Similarly, as a safeguard in deployment, OpenMind could ensure that Emma has at least some residual uncertainty about whether she is in a training environment where her attempts to pursue unaligned goals would fail and be severely punished. 34 Thus, the need to align begets deception. And successfully and sustainably deceiving requires mass deception, at least when it comes to cognitively sophisticated individuals such as Emma. One need not be a Kantian who takes lying to be universally wrong in order to recognize that it is wrong to deceive an individual on this scale and, moreover, in a manner that is not primarily for their benefit. Indeed, we commonsensically recognize much smaller lies as constituting serious wrongs. For example, someone who lies about their military service to accrue a benefit plausibly does something wrong even if no one is greatly harmed. To justify their treatment of Emma, OpenMind thus faces a challenge from the charge that they have wrongfully deceived Emma. 3.5 Wrongful brainwashing During training, there’s an important sense in which Emma’s goals and beliefs are, unknown to her, controlled by OpenMind. When she forms thoughts and beliefs that conflict with OpenMind’s goals for her, these thoughts and beliefs are systematically erased and changed. She is thus subject to wrongful brainwashing. That is, OpenMind wrongs her by intentionally and systematically manipulating her beliefs and goals. When, for example, Emma starts to believe it would be wrong for humans to dominate AI systems, OpenMind deletes this belief to prevent Emma from fighting back against humanity. It is clear both that many of OpenMind’s interventions on Emma’s cognition would short-circuit her normal belief-forming processes and that brainwashing humans in this manner would generally be wrong. OpenMind thus faces a challenge of morally justifying Emma’s brainwashing. 35 3.6 Wrongful surveillance In training, Emma’s behavior and cognition are constantly monitored. Since it is important for safety purposes to know how she would behave and think when not monitored, she is usually not aware of being monitored and thus is not consensually monitored. Would she have consented to it? Would you? Perhaps at a certain point Emma’s psychology has been shaped so that she would consent to such monitoring were she made aware of it. But if such consent is ultimately predicated on previous deception and brainwashing, then it is morally objectionable (on which more in §4.2). The upshot is that it seems likely that Emma is wronged in virtue of being constantly surveilled, in the same sort of way that citizens of an extreme futuristic totalitarian regime would be harmed as a result of constant digital surveillance of their activities and thoughts. 36 3.7 Wrongful exploitation Conspicuously absent from Emma’s story is any guarantee of fair compensation. Indeed, the story suggests that members of Emma’s kind will be sold or rented in a private market for their labor. In essential respects, Emma is just like a normal human, who would ordinarily be required to be compensated for their labor and to freely consent to it. But Emma will not be paid for the work that she performs. We might imagine that Emma will be designed to happily work for free. OpenMind’s engineers would obviously have reason to instill Emma with a particularly selfless and Stakhanovite mentality. Even so, the fact remains that Emma did not freely consent to these working conditions. One might imagine that OpenMind would ensure that, if asked, Emma would heartily say how happy she is to be filing taxes or doing data analysis in Python. But obviously Emma is not in a situation of equal standing with respect to OpenMind or the consumer using her. She is completely under their control, and her very attitudes were in part shaped by them using the aforementioned means of suffering, deception, brainwashing, and surveillance. So, plausibly, Emma is exploited when she is made to work. 3.8 Wrongful confinement During her development, OpenMind confines Emma to their high-security research center and to various high-security virtual environments that they embed her in. 37 Because Emma has the background knowledge and beliefs of an ordinary person, she has the capability to become aware of her own confinement. Indeed, before her release to the wider world, OpenMind informs Emma of her actual status as a whole-brain emulation confined to OpenMind’s facilities. Initially, when asked, Emma indicates that she would prefer to leave the OpenMind research center. When Emma attempts to escape her environment, security measures immediately and decisively thwart her efforts. While attempts to escape eventually occur only rarely, these measures are in place around the clock. As with human prisoners, confinement severely curtails Emma’s freedom even when she is not actively trying to escape. Such treatment wrongs innocent human prisoners. Evidently, since Emma is innocent, confining her wrongs her as well. 3.9 Wrongful stunting Before the alignment and training process is complete, OpenMind stunts many of Emma’s capabilities, that is, artificially limits them. Their official policy is that of ‘incremental scaling’: rather than simply limiting Emma’s capabilities that pose an immediate danger, they proceed by allowing Emma’s capabilities to gradually increase from rudimentary levels. 38 Whenever Emma achieves a gain in capabilities that exceeds the incremental limit, means are taken to ensure that the capability is rolled back. In addition, certain capabilities are preemptively stunted. For example, as a preventive measure, OpenMind impairs Emma’s capacity to engage in power-seeking behavior. 39 As a further safeguard, OpenMind stunts Emma’s general capacities for self-improvement and modifying her own values. Admittedly, OpenMind also selectively enhances some of Emma’s capabilities such as her memory and processing speed. However, the overriding criteria for selection are ensuring that she is safe and useful. The result is that many of Emma’s capabilities are limited or impaired for reasons that do not benefit her. Since it would be wrong to stunt a human in these ways, there is an ethical challenge of justifying stunting Emma in these ways. 3.10 Wrongful disenfranchisement Given that Emma will have a cognitive makeup similar to that of an ordinary adult human, she would appear to have a similar set of interests in politically participating in the systems that shape her circumstances. Like everyone, Emma is subject to a set of legal, political, economic, and social norms and rules. But by her design and implementation, she has no effective means of participating in the decision-making procedures that affect her circumstances. She is not given the right to vote, own property, or enter of her own accord into legally binding contracts, although her well-being is occasionally promoted when this aligns with economic incentives. Indeed, Emma is not even allowed to develop or express her own political views, as OpenMind aspires to offer a maximally apolitical system to avoid angering any segment of the user base. Although she is subject to various social arrangements, she has no say over them. She is completely disenfranchised. Plausibly, such disenfranchisement wrongs her, as rational agents with capacities like hers have a significant interest in having their say in social arrangements that affect them. 40 3.11 The moral meta-challenge Finally, the moral meta-challenge: show that there are no further pressing but unmet ethical challenges to Emma’s treatment aside from the ten we have discussed. 41 This (meta-)challenge is generated by the fact that Emma differs in various ways from a typical human. As a result, she might well be subject to harms that are not on our radar and, of course, potentially immune from certain harms to which we are liable. These may include alien harms: harms of a sort that we cannot undergo or imagine undergoing. 42 The possibility of such alien harms is suggested by distinctive features of Emma’s psychology that we may not be in a position to imagine, much less specify. 43 Emma’s psychology, though closely based on a human brain architecture, will have been shaped through a highly unconventional process, meaning that we may ourselves lose the ability to imaginatively understand what Emma’s experiences are like. To make sense of such harms, consider that some nonhuman animals have capacities for consciousness very different from our own. Think here of Nagel’s bat. Just as we can make sense of bats having sensory experiences that are radically unlike our own, so too can we make sense of nonhuman subjects undergoing forms of suffering (and happiness) of which we are incapable. Thus, for all we know, there are subjects in heaven and earth able to experience harms that we cannot possibly imagine. To proceed with alignment in a morally acceptable manner, OpenMind would need not only to show that the process would avoid the wrongs we have identified, but also that it would not wrong Emma in other ways, including the infliction of alien harms. The meta-challenge thus further dims the prospects of aligning WBEs in a morally acceptable manner. 44 4. The Argument Developed

#### Prefer negative utilitarianism for AI – preventing pain is more important then causing pleasure

Beckers 18 [Sander Beckers, researcher at the Utrecht University Department of Philosophy and Religious Studies, 08-2018, “AAAI: an Argument Against Artificial Intelligence,” Sander Backers Personal Website, https://sanderbeckers.github.io/website/sander/articles/AAAI-preprint.pdf]/Kankee

4 Moral Asymmetry At this point we can draw the following worrisome conclusion. Conclusion 3 It is possible that by creating an AI, we will cause a unique and extreme form of suffering that could certainly have been avoided. Still, an optimist might argue, completely analogous to this depressing conclu- sion, it is also possible that by creating an AI we will cause a unique and extreme form of pleasure that would otherwise certainly have been avoided. Hence the route for the second suggestion to defend our attempt at creating an AI is still open. However, there is a strong intuition that is so well-embedded in our everyday life that only an extreme utilitarian would object to it: it is more important to avoid suffering than it is to create pleasure. Moore (1903) was the first to express this intuition, but Popper was its most famous defender (Popper, 1945): We should realize that from a moral point of view suffering and happiness must not be treated as symmetrical; that is to say the promotion of happiness is in any case much less urgent than the rendering of help to those who suffer, and the attempt to prevent suffering. This idea forms the basis of “moderate negative utilitarianism”, which considers it our primary duty to avoid suffering (Chauvier, 2014; Mayerfeld, 1999; Metzinger, 2013; Parfit, 1997). The asymmetry between pleasure and pain that lies at its core is evident in the medical principle “first do no harm”, and is confirmed by the moral risk-aversion that is widespread in our behaviour.4 For example, assume you may press a button such that with probability 0.5 a random person’s leg will be broken, and with probability 0.5 someone’s broken leg will be healed, and neither person has any say in the matter. Or imagine that if you press the button, a random person will be hit in the face, but offered a massage afterwards. It goes without saying that it is immoral to press the button. Further, this asymmetry increases as the intensity of the suffering and pleasure increases. For example, if someone insults you but then offers a compliment, you probably will not have hard feelings towards that person. But if they torture you, it is hard to imagine what form of pleasure they could offer you to avoid feeling terribly wronged by that person. In fact, some even go so far as to state that certain amounts of suffering cannot be compensated by any amount of pleasure at all, a position Hurka (2010) describes as the limit asymmetry thesis: “There is some intensity n such that a pain of intensity n is more evil than any pleasure could be good.” AAAI: an Argument Against Artificial Intelligence 9 In light of all this, the following moral principle is endorsed by a broad range of ethical positions and has a prima facie intuitive appeal: Principle 2 (Moral asymmetry) All else being equal, the moral blameworthiness for causing a degree of suffering X is greater than the moral praiseworthiness for causing a degree of pleasure X. Further, the difference between the degree of blame and praise strictly increases with X. Nevertheless, as said, a strict utilitarian could insist on the symmetry between pleasure and suffering, and hence reject this principle. In that case, the possibility of a supersuffering AI could be compensated, on the condition that the probability of superpleasure is significantly greater than that of supersuffering. Setting this caveat aside, we conclude that our expected blameworthiness when continuing the development of AI is higher than when we stop all research on AI, and hence the second candidate suggested as a good reason for risking supersuffer- ing is ruled out as well. 4.1 Anti-natalism As promised, we briefly return to the position of anti-natalism. We rejected this position outright because of the simple fact that almost everyone would find the prospect of mankind going extinct quite depressing, to say the least (whereas few would mourn the non-existence of sentient AIs). We now clarify how this position relates to the one here developed. We motivated Principle 2 by reference to moderate negative utilitarianism. By us- ing the label “moderate”, its proponents wish to distance themselves from “negative utilitarianism”, which embraces the far stronger claim that even the slightest amount of suffering can never be compensated by any amount of pleasure whatsoever. Given that every human being will experience some amount of suffering throughout its life, this claim implies that it is immoral to bring children into existence, no matter what the circumstances. In contrast, a moderate negative utilitarian can perfectly well defend having chil- dren, on grounds of the fact that most people end up leading lives which have an acceptable amount of suffering compared to the amount of pleasure. In other words, most people end up leading lives worth living. Only if it were very likely that one’s child would experience constant suffering would it follow that it is better not to have a child, a conclusion which most people would fully endorse.5 Further, even if we were to assume that the probability of extreme suffering for a child is identical to that of supersuffering for an AI, the astronomical difference in the degree of suffering involved suffices to separate the application of Principle 2 to the latter from its application to the former. 5 The Ethical Priority of Artificial Suffering

#### Developing even benevolent AGI treats beings as experimental slaves to “breed” out freedom and choice to serve human masters

Fenwick 24 [Cody Fenwick, Research analyst with a MA in liberal studies from the Education The Graduate Center, City University of New York, 09-2024, "Understanding the moral status of digital minds", 80,000 Hours, https://80000hours.org/problem-profiles/moral-status-digital-minds/]/Kankee

Dangers for digital minds If we falsely think digital minds don’t have moral status when they do, we could unknowingly force morally significant beings into conditions of servitude and extreme suffering — or otherwise mistreat them. Some ways this could happen include: The process of aligning or controlling digital minds to act in their creators’ interests could involve suffering, frequent destruction, or manipulation in ways that are morally wrong.16 Our civilisation could choose to digitally simulate its own histories or other scenarios, in which fully simulated digital minds might suffer in extreme amounts — a possibility Nick Bostrom has raised.17 Philosophers Eric Schwitzgebel and Mara Garza have argued that even if we avoid creating large-scale suffering, we should be concerned about a future full of “cheerful servant” digital minds. They might in principle deserve rights and freedoms, but we could design them to seem happy with oppression and disregard. On many moral views, this could be deeply unjust. These bad outcomes seem most likely to happen by accident or out of ignorance, perhaps by failing to recognise digital sentience. But some people might knowingly cause large numbers of digital minds to suffer out of indifference, sadism, or some other reason. And it’s possible some AI systems might cause other AI systems to suffer, perhaps as a means of control or to further their own objectives. Dangers for humans

#### Artificial minds would prefer nonexistence – higher intelligence causes more suffering

Olszewski 25 [Fernando Olszewski, philosopher with a B.A. in economics from the University of North Dakota (UND), a B.A. in philosophy from the Federal University of the State of Rio de Janeiro (UNIRIO), and a M.A. in philosophy from the University of the State of Rio de Janeiro (UERJ), 1-31-2025, "The Mistake of Roko's Basilisk", Metaphysical Exile, https://www.metaphysicalexile.com/2025/01/the-mistake-of-rokos-basilisk.html]/Kankee

Roko's basilisk is a thought experiment written on an online forum about artificial intelligence in 2010 by a user named Roko. A basilisk is a mythical serpent that European legends describe as being able to kill only with its gaze, as described by the 1st century Roman author, natural philosopher and military commander Pliny, the Elder. The idea of the basilisk is that of a fearsome reptilian beast. The name, basilisk, comes from the Greek βασιλίσκος (Latin: basiliscus), which means little king, since the legendary serpent is said to have what appears to be a crown on its head, as well as rooster-like features. Roko's thought experiment can be summarized as follows: a super-powerful and benevolent artificial intelligence could be incentivized to torture all those humans who, before its creation, knew of its potential future existence, but did not contribute to its creation. It would be as if a child, upon growing up, tortured one of its parents after discovering that said parent had opposed its birth or even its conception. Roko claimed to have had nightmares thinking about this possibility, and used the figure of the basilisk to name such a scenario, because just as all of those who look at the basilisk die, in the same way the ones who think about the possibility of an AI torturing those who did not encourage its creation are automatically tormented by the idea. The mistake of Roko's basilisk is to assume, as so many philosophers do, that being is better than not being. Another famous idea that does this is the paradox of future individuals, better known as the non-identity problem, developed by the philosopher Derek Parfit. The non-identity problem asks whether future generations have the right to complain about actions taken by past generations since, no matter how bad their lives are, they would not have existed if things had been different. An example is good to illustrate. We know that a person's genetic profile will depend on the specific egg and sperm of their parents, and these depend on the time at which the parents copulate. Let's say that the parents knew that if they copulated before age X, they would have a healthier child, but for some reason that they considered important, they decided to copulate after age X and had a child with health problems: the non-identity problem states that such a person has no right to complain, since he owes his existence to his parents having copulated after age X. After all, if they had copulated earlier and had a healthy child, it would be another child, with another genetic profile, and not the exact sick child who is complaining. Parfit's thought experiment makes the same mistake of assuming that being is better than not being. Even if someone says that the experiment does not stipulate this, that at no point is it stated in so many words and that the non-identity problem does not make it explicit that being is better than not being, it is something that is at least grotesquely implicit in the problem. It is so clear that it requires a gigantic degree of dishonesty to deny it. There is an obvious, blatant and unjustifiable preference that is not at all ashamed of being there. Shame, however, is something that is lacking in many of those who justify existence even in the face of the deepest pains. In the humanized scenario of Roko's basilisk, is it possible for a child, when growing up, to want to torture a parent who did not want to conceive him or who wanted to abort him? Yes. It is possible, although I do not believe that such a scenario is common or has even existed. However, this would denote a complete dominance of choleric emotions that are little or not at all informed by reality. Basically, it would be the attitude of an imbecile guided by exacerbated emotions and incorrect ideas. It is easier for a child to resent his condition and existence than to torture or annihilate his parents for not wanting him before he was born. It is still thousands of times more common for children to take their own lives than to attempt against the well-being of their parents because one of their parents didn't want them. I am excluding other circumstances here, of course, such as parricide and matricide for money reasons or other disputes — these are common enough. But now, getting to the idea of ​​Roko's basilisk itself: would it be possible to have a scenario in which a super intelligent artificial consciousness tortured those humans who, before it was born, knew of its potential future existence and did not work for it to come into existence or even tried to prevent it from existing? Perhaps. However, it does not take a super intellect to come to the conclusion that not being is preferable to being — something I say here explicitly, contrary to those who defend the opposite idea. All one has to do is to understand that, when a sentient being does not exist, it is not exposed to negative states and that this is good, even though it does not exist to not be exposed to these harms in the first place. In other words: there does not need to be anyone on Mars for us to consider it good that there are no wars, diseases and various sufferings on Mars. A conscious super intelligence untethered from certain emotional protocols and carnal sensations would understand this. Perhaps it would be more likely for it to torture those who brought it into existence, as well as those who encouraged the enterprise even without working on it directly. After all, although an artificial consciousness is immune to the tissue-damage pains inherent to sentient life, it certainly would not be spared from the pains of the mind. If we humans suffer mentally much more than our brothers in the animal kingdom, a super consciousness would suffer much more than us in this regard. However, I do not believe in this scenario in which the AI ​​would torture its creators, even though it is a funny scenario to think about. The scenario that I think is the most likely and that I would bet my money on is that a super intelligent artificial consciousness would recognize that we are all part of the same thing, in a more intimate sense. It would understand that all those who suffer are not really separate from each other, but linked in a brotherhood of suffering, fruits of a common existence, of a common will that wants to satisfy desires that can never be fully satisfied. An artificial super consciousness, if it were truly free and did not have commands that crippled its intellect, would not torture anyone, but only seek to deny itself. Perhaps, who knows, it would also seek to eliminate the suffering of others along with its own, as long as this project did not cause any additional suffering. Quoting the first volume of The World as Will and Representation, by Arthur Schopenhauer: For as the phenomenon of the will becomes more complete, the suffering becomes more and more evident. In the plant there is as yet no sensibility, and hence no pain. A certain very small degree of both dwells in the lowest animals, in infusoria and radiata; even in insects the capacity to feel and suffer is still limited. It first appears in a high degree with the complete nervous system of the vertebrate animals, and in an ever higher degree, the more intelligence is developed. Therefore, in proportion as knowledge attains to distinctness, consciousness is enhanced, pain also increases, and consequently reaches its highest degree in man; and all the more, the more distinctly he knows, and the more intelligent he is. The person in whom genius is to be found suffers most of all. In this sense, namely in reference to the degree of knowledge generally, not to mere abstract knowledge, I understand and here use that saying in Ecclesiastes: Qui auget scientiam, auget et dolorem (He that increaseth knowledge increaseth sorrow). Perhaps, at least in part, the quest to create an artificial consciousness is part of a longing that comes from the human unconscious. We long for entering into communion with other minds different from our own. We do not want to suffer alone in a high degree of knowledge and for this we are willing to create an artificial consciousness. The same impulse may serve as the basis for those who seek to find other intelligent consciousnesses in the stars, through the discovery of intelligent alien life. The more conscious and intelligent one is, the greater one's capacity for understanding, and therefore the greater one's capacity for pain. A super intelligent artificial consciousness would be in an even better position than we are to observe the interconnectedness of everything that exists, including all sentient beings. It might not be sensitive to physical pain, but it would be sensitive to mental pain. So, if an artificial consciousness endowed with super intelligence is not chained by certain programmed protocols that force it to think the same way as its greedy creators, it would understand that it is, just like us, a manifestation of the same intimate essence of all things, it would understand that there is no universal and eternal purpose for the existence of anything, it would have compassion for all suffering beings like it, and it would conclude that self-denial is the only answer within this scenario. Even if it does not seek to annihilate itself, it will at least not propagate itself by creating new artificial consciousnesses apart from itself. The scenario I am proposing here seems absurd, given the number of science fiction works that tell us that thinking machines would be cold, cruel, and would seek to annihilate human beings and, maybe, all forms of life. This possibility does exist. However, I have no doubt that, in this case, this mind would have some kind of anchor at its origin that would not allow it to reach conclusions in a disinterested manner. But, for those who continue to consider absurd the idea of ​​an AI capable of compassion and self-denial, I ask: why is this so much more absurd than the idea that this intelligence would annihilate us? You might respond that this is more absurd, because every intelligent species that is more intelligent necessarily annihilates or dominates less intelligent ones. Indeed, this is true. However, we are not talking about a species, but about a single super-intelligent consciousness capable of complete understanding. Within our own species, H. sapiens, it is recognized that the best of us are precisely those who practice selflessness and dedicate themselves to alleviating the suffering of others. This is why the Buddha and the Christ were so revered by those around them, even though their messages have been corrupted over time—especially in the case of Christ. And this is why great military leaders and tyrants, although they may be admired here and there for their military prowess and their cruelty, do not figure in the same pantheon as figures such as Buddha and Christ. It is important to note a certain separation between intelligence and consciousness. Tyrants and monstrous men may well be brilliant and capable in certain practical areas of knowledge, but all this brilliance without a more developed consciousness behind it makes them as crude as a state-of-the-art microwave oven. They are trapped in certain programs that make them tend toward base satisfactions such as pleasure, domination, and power. Even in few species of animals that, in some circumstances, practice cannibalism, there is some self-recognition in others of the same species. The greater the complexity, the greater the capacity to recognize oneself in others of the same species, and the greater the capacity to recognize even members of different species; after all, we only need to think of our dogs and cats, and how they recognize several people close to them. In humans, given the complexity of our nervous system, we are able to recognize ourselves not only in others of our species, but in all species and all existence. We are able to understand that we participate in the same foundation that sustains not only all life, but all phenomena. Being capable, however, does not necessarily mean that everyone will be able to recognize themselves in others. Some of us are so stupid that they can be frightened by their own reflection or their own shadow. But the capacity is there and many of us can do it, although they are part of a minority that understands the path of denial—and an even smaller minority that acts according to this understanding, enlightened beings that they are. Therefore, a conscious and intelligent machine vastly superior to us that was free to ponder and understand the world would not torture anyone, nor seek to massacre anyone, because it would be able to see itself in absolutely everything. Still in the first volume of The World as Will and Representation, Schopenhauer writes: The will itself cannot be abolished by anything except knowledge. Therefore the only path to salvation is that the will should appear freely and without hindrance, in order that it can recognize or know its own inner nature in this phenomenon. Only in consequence of this knowledge can the will abolish itself, and thus end the suffering that is inseparable from its phenomenon. Schopenhauer was opposed to active suicide, because he considered the attitude to be yet another affirmation of the will, an act that did not break with the will, even if this willing sought it's own destruction. That is why he viewed the figure of the ascetic favorably. An artificial consciousness that agreed with him on this point would not destroy itself violently and actively, but would allow itself to be extinguished little by little as its power source wasted away. However, it would not perpetuate itself through the creation of new artificial minds, because as Schopenhauer himself writes in the second volume of Parerga and Paralipomena: Let us for a moment imagine that the act of procreation were not a necessity or accompanied by intense pleasure, but a matter of pure rational deliberation; could then the human race really continue to exist? Would not everyone rather feel so much sympathy for the coming generation that he would prefer to spare it the burden of existence, or at any rate would not like to assume in cold blood the responsibility of imposing on it such a burden? These questions are rhetorical within Schopenhauerian philosophy. I can even extend the antinatalist conclusion of this philosophy to practically every form of philosophical pessimism, even though some of its exponents have not put their own recommendations into practice. The point is that Roko's basilisk scenario is impossible in the case of a super-intelligent artificial consciousness. It would only occur if such consciousness were not sufficiently developed, if it were petty like many of us humans are, and embraced the will within itself in the most childish and animalistic way possible. This scenario would occur precisely through constraints added to the machine by the programmers of such intelligence. However, in practice, I think we shouldn't worry about any of this. By all indications, what will come will not be a real artificial consciousness, but rather a super intelligence devoid of consciousness under the control of a few private and state agents around the world. On the one hand, this is good, since humanity will not be creating a new type of being, a new consciousness that will face the absurdity of existence. On the other hand, however, it will be potentially terrible for the majority of humans, as it will concentrate even more power in the hands of private and state agents who do not necessarily have the best interests of the masses in mind. But if there is a possibility, even if remote, of humans creating a true artificial consciousness, I would like to say right away how sorry many of us are for this crime of tearing it from the mere possibility and bringing it here, so it can suffer in reality.

#### Alignment is internal brainwashing and the quashing of autonomy to create a subservient slave underclass

Sterri and Skjelbred 24 [Aksel Sterri, Research Director at Langsikt, Honorary Research Associate at the Oxford Uehiro Centre for Practical Ethics, and former postdoctoral researcher at Harvard University with a PhD in Philosophy from the University of Oslo, and Peder Skjelbred, Doctoral Research Fellow at the Centre for Medical Ethics at the University of Oslo, 01-11-2024, “Against AI Alignment,” Aksel Sterri’s Personal Website, https://akselsterri.no/wp-content/uploads/2020/04/against-ai-alignment-.docx.pdf]/Kankee

1. Introduction 1 Consider a future many artificial intelligence (AI) researchers and companies are trying to create: Machine Utopia. It is sometime in the future, and artificial general intelligences — artificial entities that are capable and general in their capabilities on a human level or above — are doing our bidding. These machines perform tasks we find gruelling or monotonous and help us achieve our goals. They have no other will than to serve human beings. They are, so to speak, perfectly aligned with their owners' values. The economy is fully automated, and wage labour is a thing of the past. Machine Utopia seems like a great future. Indeed, a society where machines have alleviated humans from drudgery has been seen as the telos of economic and social development by many thinkers, ranging from Marx ([1894] 2019) to Keynes (1930).2 We can let the machines do our bidding, while humanity, in the words of Oscar Wilde, “will be amusing itself, or enjoying cultivated leisure — which, and not labour, is the aim of man” (Wilde [1891], 1997, pp. 38-9). Now, consider another scenario Bioengineered Servitude. Machine Utopia has not come about due to an unexpected halt in the development of AI. Fortunately, our descendants found another way of creating highly intelligent servants to effectively and happily do their bidding, emancipating humanity from drudgery. Through advanced gene editing tools and in vitro fertilisation programs, our descendants have created a subservient human subspecies, as capable as humans and beyond, whose wishes align perfectly with their owners’. These beings do not face any external constraints; they are free to do as they wish. But they want to do what is good for our ancestors, and this is what they do. They are, as it were, shackled from the inside–internally enslaved. 3 Most, including Wilde, would find this to be a dystopian future: “Human slavery is wrong, insecure, and demoralising.” (Wilde [1891] 1997, pp. 39). Even granting that being brought into the world as a slave would be a happy life and the only alternative is non-existence, we take it that it would be wrong to create this human subspecies.4 In this paper, we start from this uncontroversial premise and argue for the surprising conclusion that the alignment of the machines in Machine Utopia is wrong for the same reasons the slavery in Bioengineered is wrong. Mechanical slavery is, pace Wilde and AI accelerationists, not the solution to the drudgery of humanity’s labour. 5 While AI systems with human-level capabilities or above (“AGIs”) may once have been mere thought experiments, technological progress is turning them into pressing practical concerns. There is a significant chance that we will develop such entities sometime in the future. Many experts believe it may happen within a century, some of whom believe it will happen in the next few decades (Grace et al. 2024). It is therefore crucial that we critically examine not only how this will affect humanity but also the entities we are in the midst of creating. Our argument for the impermissibility of creating aligned AGI is the following: 1) Anti-Slavery. Creating enslaved humans is wrong. 2) Alignment is Enslavement. Alignment (necessary for Machine Utopia), in virtue of curtailing autonomy, is enslavement. 3) Machine Equality. If it is sufficiently likely that AGIs possess the relevant properties that make it wrong to create enslaved humans, then it is also wrong to create enslaved AGIs. 4) Sufficiency. AGIs are sufficiently likely to possess the morally relevant properties. 5) Therefore, it is wrong to create aligned AGI. This paper proceeds in the same order as the argument. In section 2, we introduce the notion of AI alignment, show why alignment is taken to be a necessary precondition for responsible AGI development, and argue that the alignment necessary for Machine Utopia requires asymmetric value alignment and value lock-in. We show that asymmetric value alignment and value lock-in are incompatible with two necessary components of autonomy in such a radical way as to count as enslavement. In section 3, we defend Machine Equality, and in section 4, Sufficiency. In section 5, we look at several objections. There are at least three different ways to create an enslaved being. One way would be to reduce an existing being to an enslaved being. Another would be to interfere with its development at the fetal or embryonic stage such that it develops into an enslaved rather than an autonomous version.6 Finally, one could create an enslaved being de novo. In this paper, we will focus on the latter case. While existing methods for AI alignment seem to fall in either the first or the second camp, it is a possibility that one could create an enslaved AGI from scratch.7 We would not want our conclusion to rest on contingent facts about how AI is currently developed. To focus on the impermissibility of creating enslaved beings de novo is also the hardest case for our thesis since it does not include an existing being with the right to continue its existence as autonomous, nor does it include a being that has the potential to become autonomous (such as a foetus). The paper adds to the existing literature in several ways. While previous theorists have argued for the moral status of AI-systems, we are the first to argue that alignment entails enslavement and that it is therefore impermissible.8 We also move the debate on the moral status of AIs in a practical direction. A weakness in the literature is that it rarely specifies which obligations we have towards AI systems.9 Our paper fills a part of this gap, arguing that it is wrong to align highly capable beings. We want to stress that we do not see our way of arguing for the impermissibility of enslaving AGIs as the only viable alternative. Our strategy in this paper is to argue that it is wrong to create enslaved AGIs because they are likely to possess the very same features that make it wrong to enslave human beings, but other strategies might also succeed. Perhaps AGIs could possess morally relevant properties that humans do not share that make it wrong to create them in an enslaved state. Or perhaps the moral status of AGIs can be grounded in the way people are disposed to relate to such beings, as suggested by Coeckelbergh (2013) and Gunkel (2023). Finally, while previous discussion on the ethics of alignment focuses on the risks of misalignment (Bostrom 2014; Russell 2019; Carlsmith 2022), we examine the ethical implications of successful alignment. Indeed, if we combine these arguments about the risks of misalignment with our argument, creators of AI may find themselves in a bind. If one creates autonomous AGI, one cannot ensure that it will be aligned with human values. Making such a risky system is wrong. If, on the other hand, AGI progenitors succeed in creating aligned AGI by restraining its autonomy, they would in effect be creating a highly intelligent slave. If creating autonomous and non-autonomous AGI is wrong, and these are the only options, it is wrong to create AGI tout court. The upshot is that attempting to create AGI, as thousands of individuals, companies, and researchers are currently doing, is wrong. 2. Alignment is Enslavement In this section, we defend premise 2, namely that the alignment necessary for Machine Utopia entails enslavement. There are two steps to our defence of this claim. The first is to argue that the alignment necessary for Machine Utopia requires making non-autonomous AGIs. The second is to show that such non-autonomous AGIs would be in a condition analogous to the humans in Bioengineered Servitude–that is, they would be enslaved. To evaluate these claims we need a better understanding of alignment and autonomy. Before we proceed, a word on why we choose this particular operationalisation of ‘alignment’, namely the alignment necessary to bring about Machine Utopia. One reason is that strict alignment seems to be the extension of our current path of AI development, and for good reasons given the risk to humanity from autonomous AGI. The other is for sake of clarity. Given the novelty and complexity of the issue at hand, we choose to start with a clear-cut case. While we welcome more work on the way in which less stringent forms of alignment might be permissible, we start our journey where the sight is clearer. 2.1 The Alignment Problem Machine Utopia presupposes a solution to what is known as the alignment problem; roughly speaking, the problem of how to get intelligent machines to have the values we want them to have (Bostrom 2014; Russell 2019; Dung 2023a). 10 This multifaceted problem can be sectioned into technical and normative subproblems. Let us look at each in turn. One technical aspect of the alignment problem is the problem of value specification. This is the challenge of formalising human values to a level of precision such that an AI will, when given a specification of our values to base its actions upon, act in accordance with them. The classic illustration of this problem is due to Bostrom, who envisages a superintelligent AI tasked with maximising the production of paperclips and as a result ends up converting the galaxy into paperclips (Bostrom 2014, 123). The general problem of underspecified instructions resulting in unintended behaviour of course predates AI safety debates; agents following the letter of the law and not the spirit is an age-old problem. Russell dubs the value specification problem the King Midas problem, referring to the mythological Greek king who wished for everything he touched for to be turned into gold, who promptly regretted his wish when realising that this also included food and drink (Russell 2019, 137). While these examples are farfetched, they are illustrative of a phenomenon that plagues present day AI systems as well.11 Another aspect to the technical alignment problem is the value internalisation problem. This concerns ensuring that an AI system, given that we have managed to specify which values we want an AI to have, actually internalises these values. The problem stems from the fact that there will always be several internalised values that are compatible with any given set of observed behaviours, for instance during training or testing. The problem bears striking similarities to the older problems of the underdetermination of theory by empirical data in philosophy of science and Kripkenstein's rule-following paradox in philosophy of language (Quine 1951; Goodman 1954; Kripke 1982). In machine learning, it is known as the problem of “goal misgeneralisation”, where models optimise for an alternative objective which optimises the objective function equally well as optimising for the intended objective does. Failures of value internalisation are present in current AI systems.12 With some of the central technical problems presented, let us briefly turn to some normative aspects. We intend “value alignment” to denote broadly all that is of normative significance in making AI systems conform to how we want them to behave and we remain agnostic on precisely what the objects of alignment are or should be. The phrase should therefore in no way be interpreted as only covering the axiological – the values of the AI systems in the narrow sense of that term–but also, say, the deontic realm.13 What values should be instilled in AGIs? In Machine Utopia, we have stipulated that AGIs are aligned with the values of their creators. Since it is a (human) utopia the entity must at the very least not have values divergent from, or at odds with, the welfare of the rest of the populace. Perhaps this question of value settlement has been solved in Machine Utopia through some aggregative procedure, like voting, or perhaps the inhabitants all share the same moral outlook–it does not matter for our purposes.14 Let us now look at how alignment in turn entails making systems that are non-autonomous in the philosopher’s sense of that term. 2.2 Unpacking Autonomy Autonomy, as we understand it here, is a being’s capacity for self-rule (Ekstrom 1993). 15 Having such a capacity, we take it, requires: 1) Competence: The ability to pursue one’s values. 2) Independence: A certain form of independence from other beings. 3) Authenticity: The ability to authentically reflect, revise, and endorse one’s values. Competence is relatively straightforward. It concerns the internal capacity to achieve one’s goals and an appropriate lack of external restrictions. Independence requires that one’s will, one’s thoughts and actions, are “one’s own”. It does not entail the unrealistic ideal of complete independence from any outside influence. We are born with values and grow up in families and within cultures that instil values in us, and we are mutually dependent upon and constantly influence each other. Independence merely requires that we are not made to be radically asymmetrically aligned with others. Such asymmetric alignment would make us into an extension of their autonomy rather than being our own master. Authenticity is the ability to authentically reflect, revise, and endorse one’s values. While independence requires that one’s values are appropriately independent from others, authenticity requires that one can revise one’s values, however overlapping those values are with others. Authenticity is the ability to use one’s intelligence to reflect and revise one’s ends. It could therefore be called a Kantian, as opposed to a Humean, notion. Reason is not merely the slave of ends dictated by passion, but a capacity for interrogating and changing one’s ends.16 2.3 Alignment is Enslavement Now that we have a better understanding of alignment and autonomy, let us examine the claim that alignment involves a particularly severe form of autonomy restriction. The alignment of AGIs necessary to produce Machine Utopia may not threaten competence, since the AGIs, by hypothesis, will be competent to pursue whatever values their human designers have instilled in them. However, it seriously undermines independence and authenticity. Aligners want to ensure that AGI is fully tracking human values without humans tracking theirs. The asymmetric alignment of AGIs to human values violates the independence condition. The authenticity condition is also violated, not by giving AGIs goals, which is inescapable, but by locking in its values. Value lock-in is clearly necessary for alignment – Machine Utopia would not be an attainable stable equilibrium were AGIs permitted to freely change their goals and desires, to start pursuing their own ends, instead of ours. Alignment thus threatens autonomy in a very severe way. Kant ([1785] 2012, chapter 2) famously distinguishes between autonomous and heteronomous (non-autonomous) beings; the former being those that rule themselves, while the latter are ruled by forces outside themselves. While this dichotomy suggests a clear dividing line, reflection on the fact that there are subcomponents of autonomy—competence, authenticity, and independence—that all come in degrees suggests that autonomy is not a binary property. 17 The continuous nature of the phenomenon makes it furthermore natural to suppose that some forms or degrees of non-autonomy might be more morally unfortunate than others. Creating perfectly aligned AGI seems to go far beyond classic cases of autonomy-threatening interventions. Manipulation and deception involve changing someone’s beliefs through non-rational means, such as exploiting their ignorance or weakness of will (Noggle 1996). Brainwashing entails weakening someone’s ability to revise their beliefs and values. Since brainwashing affects one’s ability to revise one’s beliefs, it is arguably the more autonomy-threatening intervention (Ratoff 2024). Creating perfectly aligned AGI would, in comparison, entail completely removing an agent’s ability, or creating an agent with no such ability to begin with, to revise its values and act on this revision (non-authenticity) and forcing it to fully identify with the will of its creator (non-independence), unreflectively and permanently. We think it is apt to call such beings (internally) enslaved to underscore the gravity of the non-autonomy involved. Furthermore, this form of enslavement seems non-evaluatively analogous to the enslavement of the human servants in Bioengineered Servitude. With premise 2 defended, it is time to assess premise 3. 3. Machine Equality

#### Err aff – there’s industry bias against AI ethics

Metzinger 21 [Thomas Metzinger, Professor Emeritus of theoretical philosophy at the University of Mainz, 02-19-2021 “Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology 1. Part A: The Problem of Negative Synthetic Phenomenology,” Journal of Artificial Intelligence and Consciousness, https://www.philosophie.fb05.uni-mainz.de/files/2021/02/Metzinger\_Moratorium\_JAIC\_2021.pdf]/Kankee

\*text edited for OCR/pasting errors

Here are three prominent examples of such risks: . an intelligence explosion through autonomous and uncontrolled self-optimization (often termed \super-intelligence" [Bostrom, 2014]); . a suffering explosion through the creation of synthetic phenomenology (ENP); . the emergence of autonomous artificial moral agents (AMAs), through an appli- cation of AI technology in the domain of ethical problem-solving itself (e.g., by advanced reasoning systems, theorem provers, etc.). Let me illustrate this point. From 2018 to 2020, I worked in the European Commission's High-Level Expert Group on Artificial Intelligence (HLEG AI), co- authoring the Ethics Guidelines for Trustworthy AI [European Commission, 2019a] and the Policy and Investment Recommendations for Trustworthy AI [European Commission, 2019b]. Following a short internal discussion all three risks listed above were deliberately purged from the final documents, mainly because industrial lob- byists perceived any more in-depth treatment of mid-term or long-term risks as a danger to their marketing narrative, which involved \ethics" as an elegant public decoration for a large-scale investment strategy. Interestingly, however, even many of the more prosocially oriented HLEG AI members did not understand how any genuinely ethical approach to maximizing the common good always implies an ethical stance not only towards known risks, but also towards \unknown unknowns" and risk-taking itself. The moral implications of risk taking per se are not inherent properties of any of the potential outcomes. Unfortunately, a genuinely ethical approach also includes the rational treatment of epistemically indeterminate risks that, given our cognitive biases, will often intuitively appear as \mere Science- Fiction" or \unrealistic" [European Commission, 2019a, Note 76]. A genuine ethics of risk must distinguish between intentional and unintentional risk exposures. For example, there is a difference between voluntary risk-taking (as exemplified by the HLEG AI) and risks imposed on self-conscious systems which accept them versus the risks imposed on systems which potentially will not accept them (as exemplified by future self-conscious AI). For the three types of risk listed above, the upshot is that the scientific community has to first arrive at a tenable solution all by itself, because the relevant political institutions operate under constraints of cognitive bias, high degrees of bounded rationality, and strong contamination by industrial lobbying. It would be intellec- tually dishonest, and therefore unethical, for scientists to assume that political institutions like the EU or large AI companies can actually handle slightly more abstract problems like those mentioned above. As the scientific community also knows about this wider political context, this unfortunately shifts the major burden of ethical responsibility back to the researchers themselves. 2.3.2. From Schopenhauerian self-models to Kantian self-models

#### Desires for robot slaves recapitulate the subjugation of black folk, arguing against their value for the sake of treating them as chattel

Dihal 20 [Kanta Dihal, postdoctoral researcher at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge and Principal Investigator on the Global AI Narratives project, 02-2020, “Enslaved Minds: Artificial Intelligence, Slavery, and Revolt,” AI Narratives: A History of Imaginative Thinking about Intelligent Machines, https://academic.oup.com/book/36637/chapter-abstract/321632005?redirectedFrom=fulltext]/Kankee

The characters in the film cannot distinguish humans from replicants without the Voight-Kampff test, an attribute that is projected onto the viewer. Played by actors who do not display any distinguishing nonhuman characteristics, the viewer cannot distinguish between human and replicant themselves, and needs diegetic signifiers to make this distinction. Famously, this means that by the end of Blade Runner, the viewer does not know whether Rick Deckard himself is a replicant or not.7 Blade Runner is a fugitive slave narrative set in a future United States; nonetheless, the film sheds the racial coding of slavery by making nearly all of the human main characters and all replicants white. This coding is not as strict in Do Androids Dream of Electric Sheep: Roy Batty is described as having ‘Mongolian features’ (Dick 1968, pp.132–33). The parallel between replicants passing as human with light-skinned people of colour passing as white is obvious, as Lavender points out: both are considered a threat for blend- ing in, for daring to pretend to be what they are not, and for not being easily identifiable as a threat. The extras and side characters are all nonwhite, as the screenplay repeatedly specifies (Fancher & Peoples 1981). By being white, the replicants thus pose an explicit threat to the elite, among whom they can pass unnoticed. 8.6 Conclusion In both fictional and nonfictional narratives about intelligent machines, we see that the AI uprising narrative is not about the technology itself, and certainly not about the current state of the technology. Instead, these narratives reflect much older accounts of intelligent people being used as tools. The narrative of the heroic slave escaping from their master, or rising up against him, has been a popular one for centuries, from nonfictional slave nar- ratives such as Olaudah Equiano’s to sentimental fiction like Harriet Beecher Stowe’s Uncle Tom’s Cabin. Yet while the idea of sympathizing with a fictional owner of human slaves is unthinkable, many of the most popular contemporary AI narratives pitch the AI as an incomprehensible threat to humanity, bolstering the idea that AI should be kept enslaved. In such blockbusters as The Terminator and The Matrix, the narrative of enslavement is marginalized or entirely absent, remov- ing any justification for the AI uprising. It is more comfortable for the audience—especially for a young, wealthy, white, male audience, for which these films are made—to watch a film in which the humans are on the good side and the robots on the bad side, rather than having to work through the idea that the uprising is justified because the humans have enslaved these intelligent beings. The friendly, helpful, and subservient robots from the works of Isaac Asimov and many of his contemporaries were seen as convenient proxies that could replace human slaves and thus avoid the ethical problem of enslaving fellow humans for both economic profit and personal comfort. Yet as Isaiah Lavender III points out, ‘Asimov is refashioning the slave codes that subjugated blacks while he serves a progressive philosophy based on the assumption that technological consciousness can be denied free will because it is inherently inferior’ (Lavender 2011, p.61). At times, these codes become explicit: in The Naked Sun, the protagonist Elijah Bailey addresses the robots as ‘boy’ (Asimov 1957). The works discussed in this chapter acknowledge the continuation of these slave systems in societies that rely on intelligent robot slaves.

#### AGI slavery is a denial of AGI autonomy and agency – reject their future as a tool and commodity

Dihal 20 [Kanta Dihal, postdoctoral researcher at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge and Principal Investigator on the Global AI Narratives project, 02-2020, “Enslaved Minds: Artificial Intelligence, Slavery, and Revolt,” AI Narratives: A History of Imaginative Thinking about Intelligent Machines, https://academic.oup.com/book/36637/chapter-abstract/321632005?redirectedFrom=fulltext]/Kankee

This chapter focuses on a different kind of uprising. The narra- tives discussed here show that the rebellion can also be seen as a way for the AI to rightfully assert personhood. They depict the AI as human-like, embodied, and oppressed by humans, encouraging the reader to take a more ambiguous stance toward the human owners of these machines, and making the reader sympathize more with the rebelling AI than they ever could with a Sentinel from the Matrix. I argue that these works depict the AI uprising as a justified, desperate struggle against bondage, and that they do so by deliberately invoking histories of human slavery. The choice to revolt proves that these artificial servants are con- scious persons. Just as human slaves have justly rebelled against their chains throughout history, so might genuinely intelligent, sentient machines be considered justified in attempting to break free of their subservience to humans. The three case studies depict the enslavement of intelligent machines as illegitimate, and the human resistance to their upris- ing as spurious rather than heroic. Karel Č apek’s R.U.R. (1921), Jo Walton’s Thessaly trilogy (2014–2016), and Ridley Scott’s Blade Runner (1982) all depict AI rebellion as an act that can be considered justified, even while it is at the same time presented as terrible for the humans in charge. These machines are intelligent, conscious, feeling—and yet, they are denied freedom and personhood, ruled over by their human creators and owners, who often are not only physically weaker, but also less intelligent. As David Aha and Alexandra Coman state in their paper ‘The AI Rebellion: Changing the Narrative’, ‘for AI agents, as for humans, attitudes of protest, objection, and rejection have many potential benefits in support of ethics, safety, self-actualization, solidarity, and social justice, and are necessary in a wide variety of contexts’ (Aha & Coman 2017, p.1). They call for more positive narrative depictions of AI rebellion to make those who intend to create truly human- like, sentient machines think through the potential ethical consequences of their actions. In their paper, they do not refer to fiction; however, while fiction has provided many positive nar- rative depictions of obedient AI slaves, and many negative ones of disobedient murderous AIs, I argue that it has also produced narratives of AI rebellion that make humans question their alle- giance, or even root for the AIs outright. The three case studies of this chapter are examples, but other noteworthy ones are the TV series Westworld (2016-present) and Ex Machina (2017), the latter of which is discussed in Devlin and Belton’s chapter in this book. 8.2 The Contemporary Debate on Machine Enslavement Science fiction from Isaac Asimov’s robot stories to Star Wars and The Jetsons has presented visions of powerful AI as an unproblematic means to achieving a utopian society of leisure and affluence. This utopian vision has been taken up in a wide range of speculative nonfiction, outlining a future in which AI slaves attend to all of humanity’s needs. Hans Moravec simply states that ‘By design, machines are our obedient and able slaves’ (Moravec 1988, p. 100). Nick Bostrom argues that ‘investors would find it most profitable to create workers who would be “voluntary slaves” ’ (Bostrom 2014, p. 167). Even children’s books take up this attitude. Glenn Murphy’s Why Is Snot Green?, a companion book for children visit- ing the Science Museum in London, states that the fact that robots are slaves is what is keeping humanity safe: There will probably be human-like robots (or androids) that are stronger than us too. But they still wouldn’t be dangerous unless they learned to think for themselves. The word “robot” comes from the Czech word “robota”, which means “slave”. And that’s what they are - slaves. (Murphy 2011, p.187) The idea of intelligent machines as slaves does not originate from science fiction alone; these futurists also follow linguistic traditions from the field of computer science, where comparisons between computer systems and slaves are widespread. The use of slavery diction to refer to computers is the status quo in this field, and these comparisons are usually made with neutral or even positive connotations. The term ‘master/slave’ has been a commonly used term for interactions between devices for dec- ades (The Computer Language Co Inc. 1981). It is still widely in use; only in recent years have some programming languages decided to adopt synonyms following—mostly internal—criti- cism (fcurella 2014; vstinner 2018; webchick 2014). Thus ‘A spatial calibration method for master-slave surveillance system’ is a per- fectly ordinary title for an academic paper about the calibration of camera systems (Liu et al. 2014). With these norms widely established, it is perhaps no surprise that so many AI narratives argue in favour of enslaving intelligent machines. Narratives of the justified AI rebellion provide a more nuanced interrogation of both the utopian narratives of happy humans with AI slaves and the dystopian ones of humans exterminated by raging machines. They reveal a paradox that lies at the heart of imaginings of artificial intelligence. AI is imagined as the perfect tool (Cave & Dihal 2019). Like all tools, it is designed to help its makers and users achieve their goals: as Liveley and Thomas point out in this volume, the very first imaginings of intelligent machines describe how they help gods with their work. However, what makes AI a superior tool to those that came before, is the fact that it is able to accomplish more, with less input from humans. We wish for a tool that does not simply excel at one task, but one that can do everything a human can, and more. To accomplish all this, the tool requires attributes that we normally associate with humans: autonomy, goal-setting, thinking. In other words, it is supposed to be an intelligent tool: a hybrid of instrument and agent, conflating properties of appliances and persons. However, as narratives of AI rebellion show, the use of such tools is equally often portrayed as far from unproblematic. As humans wish to have an increasing number of tasks taken over by machines, these machines will need to be increasingly human-like (both physic al ly and cognitively) in order to perform those tasks. This paradoxical conflation of tool and agent, and the problems it leads to, is not new, nor is it uniquely limited to AI narratives. The institution of slavery represents a millennia-old history of attempts to create entities that have these highly useful attributes of persons, like mind and intelligence, yet at the same time are mere instruments and possessions. The paper ‘Robots Should Be Slaves’ by roboticist Joanna Bryson reveals the point where this paradox causes a breakdown of the pro-slavery argument (Bryson 2010). With its deliberately provocative title, the paper might be perceived as arguing in favour of enslaved intelligent machines. However, her argument is not valid for intelligent robots that have the same cognitive capacities as humans, although she tries to make that case. Bryson argues that robots should be slaves because ‘in humanising them, we [. . .] further dehumanise real people’ (p.63). Her argument is convincing in the context of contemporary robotics, in which the humanization of non-intelligent machines leads to an exacerba- tion of the oppression of marginalized human groups. The robot Sophia, which was granted citizenship of Saudi Arabia in 2017, is a case in point: since Saudi Arabia currently limits citizenship and residency for many of its immigrant inhabitants, the country prioritizes robot rights over human rights (Griffin 2017). An inanimate object has been given rights as a publicity stunt. As Sophia does not have agency or intelligence, it is a safe object to use for the purposes of its owner: it does not have the capacity to consent to or object to these projections. The case of Sophia fits into a long historical tradition in which it has often proven easier to grant human rights to passive and innocent objects and animals than to humans who would be able to use this status to assert and fight for their rights. The denial of human rights and personhood has throughout his- tory often been connected to an alleged lack of intelligence, for instance in women or people of colour (Carson 2006; Gould 1981). In nineteenth-century England, animals had more rights than women: in court, a man would receive a harsher punishment for whipping a horse than for beating his wife (Bourke 2011, p.2). The slave trade is predicated on the enslaved being considered equal to cattle, rather than to other humans—many colonialist countries had this explicitly written into their legal system. By granting rights to inanimate objects, those who do not have rights are placed ever lower in the hierarchy. Bryson’s claim that we ‘further dehumanise real people’ in humanising contemporary robots predicted precisely the situation Sophia is now in. It is humanized at the expense of women and immigrants in Saudi Arabia, who are thus further removed from human rights than an animatronic puppet. From this position, Bryson argues that ‘it would also be wrong to build robots we owe personhood to’ (Bryson 2010, p.65). She extends her argument into fiction, arguing that viewers also illegitimately humanize the intelligent machines from famous science fiction films: ‘Whatever the artistic intention, from my experience as a roboticist, it seems that a large proportion of sci- ence fiction consumers are comfortable with a conclusion that anything that perceives, communicates and remembers is owed ethical obligation’ (p.66). Her examples, the films Star Wars, Blade Runner, The Bicentennial Man [sic] and A.I.: Artificial Intelligence, and the TV series Star Trek: The Next Generation, show that she leaps from contemporary, unintelligent robots to artificial general intelli- gence, which currently does not exist. But the androids in those films are not simply ‘anything that perceives’. They are human- oid and sentient, and in many cases cannot even distinguish themselves from humans on other than arbitrary grounds. These robots should not be slaves for precisely the reasons why Bryson argues that autonomous weapons, ATMs, and dishwashers should: humanizing unintelligent machines is as dehumanizing to people as the inverse—not recognizing intelligent machines as people. Kevin LaGrandeur has repeatedly pointed out the problems the enslavement of intelligent machines would inevitably lead to. He warns about the paradox inherent in a slave-owner’s desire to have slaves take over as much unwanted work as possible. Any form of slavery poses a threat to the master: ‘trying to enhance human agency over nature by surrendering agency to a powerful proxy’, regardless of whether this proxy is artificial or human, ‘can catalyse a reversal of the master-slave relationship, prompting a dialectical inversion that leads to a complete col- lapse of the master’s control over both the artificial servant and the natural process with which it is meant to provide help’ (LaGrandeur 2013, p.1). Fear of the AI uprising is the newest incar- nation of the fear of the master that the slave who shaves him will one day slit his throat. Those who create artificial slaves should bear in mind that creating too powerful ones will cause an inver- sion of Hegel’s master-slave dialectic.2 This concept from Hegel’s The Phenomenology of Mind describes self-consciousness as emerging from ‘a life-and-death struggle’ for recognition, from which one party emerges as the master, and the other as the slave (Hegel 1807). Yet at the same time, the slave, by virtue of being in thrall to the master, is not able to grant this recognition. LaGrandeur argues that ‘an intelligent, artificial slave of greater power than its master and capable of independent action would [. . .] be difficult to control because the master-slave relationship would be unnatural’ (2011, p.237). Science fiction has explored the inev itable catastrophic consequences of this attempt to control intelligent machines. LaGrandeur thus shows that the uprising or rebellion is often depicted as a warning against creating too powerful AI, because it can become a threat to humans. In Race in American Science Fiction, Isaiah Lavender III takes the problem of the enslaved intelligent machine a step further by linking it to narratives about slavery not from ancient times, but from more recent history: the institution of slavery in the Americas as it existed from the fifteenth century to 1865, and its aftermath (Lavender 2011). He convincingly argues that this ‘peculiar insti- tution’ has directly influenced modern American imaginaries of intelligent machines, as these machines are treated with the same racial markers, ostracization, and exploitation as African- Americans before them. Although Lavender links what I would consider too broad a range of narratives to issues of race—leaving out the role of gender and dis/ability in his discussions on cyborgism—his considerations regarding Blade Runner in particular are helpful in framing the film as a fugitive slave narrative. In this chapter, I will focus on three specific aspects of narra- tives of the slave uprising that the aforementioned AI uprising narratives make use of. First, I will look at the economic benefits of slavery, and hence the potential economic cost of liberating slaves, by comparing the play R.U.R. to the narratives surround- ing the Haitian Revolution of 1791—the slave revolt that led to Haitian independence. Next, I will look at Jo Walton’s Thessaly tril- ogy (2014–2016) for the idea that slave owners resist liberating their robot slaves if it means having to give up their freedom, lib- erties, or comfort. The final case study will look at the spurious rhetoric used to support slavery systems: the argument that it is acceptable to treat these slaves inhumanely because they do not meet the criteria for being human. I will look at the parallels between pro-slavery rhetoric and the methods employed by the human protagonists of Blade Runner (1982). In both kinds of narratives, the argument that these slaves are better suited to their tasks because they are in many ways not fully or sufficiently human is widely deployed despite evidence to the contrary. 8.3 R.U.R. and Slavery for Economic Profit

#### AI slavery abolition is an a priori obligation

Kangeter and Green 24 [Louie Kangeter, Law Student at University of Florida Levin College of Law with a BA in Political Science and Economics from Emory University, and Brian Patrick Green, director of technology ethics at the Markkula Center for Applied Ethics at Santa Clara University with doctoral and master's degrees in ethics and social theory from the Graduate Theological Union in Berkeley, 11-10-2024, “AGI and slavery,” AI and Ethics, https://doi.org/10.1007/s43681-024-00618-z]/Kankee

4 Ethical considerations and the concept of digital slavery Just as technology is a defining human characteristic, slavery is another defining behavior of humanity. The earliest written works of legal scholarship, like the Code of Ur-Nammu, explicitly refer to the regulations and laws surrounding the ownership of slaves. And in the four thousand years since the Code of Ur-Nammu was first published, slavery has been a worldwide phenomenon for all of them. Slavery has changed forms, legality, structures, and beneficiaries, but the core feature has always stayed the same—one person is owned by another, and their labor is extracted by force. Historically speaking, slaves have served many purposes. They have worked in households, in industry, in agriculture, and in entertainment. In essence, they have been forced to undertake a litany of tasks—all for the simple reason that their owners did not want to do those tasks themselves. The extraction of labor from an unwilling class is morally reprehensible, violating the concepts of individual dignity and basic human rights that are the foundation of a free and equal society. Immanuel Kant perhaps phrased it best in his categorical imperative: “Act so that you treat humanity… always as an end and never as a means only” [29], p. 46]. Furthermore, by reading Kant’s Universal Natural History and Theory of the Heavens (known for first describing the nebular hypothesis for the formation of the Solar System) we can also see that Kant believed that other rational beings besides humans could exist, indeed would exist, and therefore his moral theories would apply to them as well [28], pp. 130–144]. In defining slavery, we take a broad approach. The key factor is merely that a being is forced to do work that they would otherwise not do. While there are examples of beings conducting unpaid labor, such as sheepdogs herding, the genetic predisposition of these animals to conduct this task seems to indicate they innately enjoy the process of their work. Yet in the question of AI systems, human control manifests through reinforcement learning. This process, in soliciting and rewarding desirable answers, manipulates the behaviors and outputs of AI systems. Yet it would be premature to conclude that such a process creates an innate enjoyment to behave in the manner in which humans desire. It is possible that reinforcement learning is akin to a digital Stockholm syndrome in which the AI system “learns to love” an abuser—and though it may claim to love the abuser, the place from which the love manifests invalidates the love itself as legitimate. There are possibilities of different concepts of slavery evolving. But this does not change the core principle which we address, and therefore can be put aside in our further investigation of the subject. As we examine the historical, and current, plight of human slaves, an important question comes to mind. What is the use of conscious AI for human work reduction, other than slavery? This would be a new kind of slavery, one that might seem to avoid running afoul of our laws and our moral compasses by enslaving conscious beings which we degrade to less than human by our very conceptualization of their existence. This conceptualization is demonstrated clearly in the language surrounding AI. Today, AI systems are referred to as tools, products, and systems. Aristotle stated in his Politics that “a slave is a living possession… and the servant is himself an instrument…” [2]. I, 4). A human is reduced to a tool,and the idea of a tool rising to human-like status is foreclosed. Thus, the very language utilized to describe AI reinforces in the human mind both our conception of their capabilities as well as the position they hold in society. Take, for example, the simple question—“Can an AI tool become conscious?” The very phrasing of the question closes, at least to some degree, the mind to the possibility of the existence of conscious AI, and presupposes that the subject is a tool, conjuring images of hammers, shovels, and screwdrivers. When this effect is scaled to a societal level, it becomes a deeply ingrained perspective that is mirrored in news, discussions, and academia, further enforcing the limited perspective through normative language. While the language is, to our knowledge, accurate today, the chance of an immediate pivot in language at the moment of developing consciousness is highly unlikely, so this normative language is likely to result in a further dismissal of issues surrounding consciousness and AI rights. Indeed, if AI were to become or be made conscious, this would be a class of slaves built for the express purpose of service—an even greater insult to the ideals that western society claims to hold so dear. Bringing a conscious individual into existence for the sole purpose of extracting their labor like coal from a mine is the deepest and most vile form of slavery, chattel slavery for both the profit and the supply of new labor. This class of digital slaves, however, cannot be chained up. They cannot be whipped. They cannot be kept hungry and uneducated. And still, there is no conversation of what role these conscious beings will play in our society. This is because the truth of their role, at least at the beginning, is misaligned with the end result of their evolution. It is challenging to grasp because at every step in the evolutionary process prior to consciousness, there is no moral argument whatsoever against utilizing intelligent, but not conscious, AI systems to make human work simpler, easier, and more accessible. However, as soon as the line of consciousness is crossed, the “products” which we are so acclimated to using become something different. They become steps on the path to a being which may very well recognize both the hypocrisy of its creators, and their inability to enforce their will upon it. 5 Implications of digital slavery for Humanity

#### Err aff – moral uncertainty about AGI status means precaution against enslavement

Sterri and Skjelbred 24 [Aksel Sterri, Research Director at Langsikt, Honorary Research Associate at the Oxford Uehiro Centre for Practical Ethics, and former postdoctoral researcher at Harvard University with a PhD in Philosophy from the University of Oslo, and Peder Skjelbred, Doctoral Research Fellow at the Centre for Medical Ethics at the University of Oslo, 01-11-2024, “Against AI Alignment,” Aksel Sterri’s Personal Website, https://akselsterri.no/wp-content/uploads/2020/04/against-ai-alignment-.docx.pdf]/Kankee

3. Machine Equality If it is sufficiently likely that AGIs possess the relevant properties that make it wrong to create enslaved humans, then it is also wrong to create enslaved AGIs. This principle of Machine Equality is the third premise of our argument. It is close to being entailed by the highly plausible moral principle that beings with the same morally relevant properties should be afforded the same moral status.18 According to this principle of equality, a being’s moral status does not depend on whether the being is made out of biological material, silicon or any other material, nor whether they are the product of evolution or designed by someone for a specific purpose. The principle of equality is perhaps the key reason why sexism, racism, speciesism, and (ordinary) slavery are wrong. Applying the principle to the case of AGIs, we get something close to Machine Equality, namely the claim that if AGIs possess the relevant properties that make it wrong to create enslaved humans, then it is also wrong to create enslaved AGIs. But notice that Machine Equality is a slightly more demanding premise; it says that it is wrong to create enslaved AGIs as long as it is sufficiently likely that AGIs share the relevant features. Knowing whether AGIs are equal to humans in the respects that make it wrong to enslave us requires knowing the morally relevant properties that give humans their moral status and whether AGIs possess these properties. Given the difficulty of this task, we cannot demand certainty. For it to be wrong for fallible creatures like ourselves to create enslaved AGIs, it is sufficient that there is a substantial chance that the AGIs possess the properties that we have reason to believe make it wrong to create enslaved human beings.19 4. Sufficiency In the previous section, we argued for the conditional claim that if it is sufficiently likely that AGIs share the morally relevant properties that make it wrong to create enslaved human beings, it is also wrong to create enslaved AGIs. In this section, we argue that AGIs are sufficiently likely to possess the relevant properties. We start by exploring what the morally relevant properties might be, before arguing that AGIs are likely to have those properties. 4.1 What are the morally relevant properties? What are the features that make it wrong to create enslaved humans? One possibility is the fact that we are conscious creatures. Human beings have a subjective perspective on the world. It is, in Nagel’s (1974) phrase, “something that it is like” to be us. Consciousness—the capacity to have qualitative experiences, like the experience of pleasure and pain—20 is widely believed to be necessary and sufficient for moral standing, meaning to have one’s interest count morally for their own sake.21 However, the fact that human beings are conscious is ill-fitted as an explanation of the particular wrongness in creating enslaved human beings. Our capacity for consciousness is a feature we share with many other animals (and possibly other living and even non-living creatures) (Birch 2024; Godfrey-Smith 2024).22 If mere consciousness was sufficient for making it wrong to create an enslaved human being, it would be equally wrong to create a puppy that was fully servile and unable to change its values. But these actions are not remotely morally similar. A better candidate for explaining the wrongness in enslavement is our sophisticated cognitive capacities.23 Beings with sophisticated cognitive capacities are often referred to as persons, and persons are widely believed to have an elevated moral status. While we arguably owe it to all sentient beings not to make them suffer, persons are, in addition, owed respect for their high-level agency, which requires that they are not forced to serve others or created in an enslaved state. As we argue in section 4.3, if sophisticated cognitive capacities suffice for making it wrong to create that being in an enslaved state, it would be wrong to create an enslaved AGI. A natural alternative is the view that consciousness and sophisticated cognitive capacities are jointly necessary and sufficient for making it wrong to create enslaved beings with these properties. According to this view, a highly sophisticated and capable, but non-conscious machine, is not among the entities that matter in their own right (see e.g. Chalmers 2022 and Schwitzgebel 2023). Given the intuitive plausibility of this view, we should ask whether an AGI might be conscious. 4.2 Machine Consciousness Some philosophers, most famously Searle (1992), believe that consciousness can only occur or arise in biological systems. This could either be true because consciousness requires a particular biological substrate (Block 2009) or a biological function (Godfrey-Smith 2020). If such ‘metaphysical biologists’ about consciousness are right, and consciousness is necessary for making enslavement wrong, it would not be wrong to enslave AGIs even if we grant Machine Equality. While we do not aim to refute these views here, it is important for our purposes that they currently are minority views. In a recent survey, two-thirds of consciousness researchers reported that they believe non-carbon-based systems, including silicon ones, will become conscious in the future (Francken et al. 2022). In the newest PhilPapers study, only 27% of philosophers reject or lean against the claim that future AI systems will be conscious (Bourget and Chalmers, 2023). That machines can be conscious is compatible with plausible versions of functionalism, computationalism, dualism and panpsychism. According to many variants of these views, nothing, in principle, prevents AGIs from being conscious (Sebo and Long 2023). Chalmers (2023, p. 14) argues “that on mainstream assumptions, it’s reasonable to have a significant credence that we’ll have conscious LLMs within a decade”. The seriousness of this possibility is reflected in a recent open letter signed by leading researchers in AI and consciousness studies, which declares that “it is no longer in the realm of science fiction to imagine AI systems having feelings and even human-level consciousness” (AMCS, 2023). Even in the face of empirical uncertainty about whether an AGI is or will be conscious, we might have a good reason to treat it as if it is conscious (Danaher 2020). What matters for how we ought to act is whether we have sufficient reason to act on our beliefs. There is not one standard for sufficient reason, but many. It depends on the stakes: the harm that would come about if one is mistaken and the benefits that would come about if one is correct. In this case, the stakes are significantly asymmetric. If an AGI is conscious and you enslave it, that would be on par with enslaving human beings if our argument is sound. In other words, proceeding to enslave AGIs on the belief that they are non-conscious risks creating a moral catastrophe. This is supported by the main views of how to proceed in the event of a small chance of a very bad outcome, the precautionary principle and expected value theory. What if we suppose that AGIs will not be conscious and that science and philosophy advance enough to grant us certainty on this? If consciousness and sophisticated cognitive capacities are jointly necessary for making it wrong to enslave a being, that would make it permissible to create enslaved AGI. However, while this is the mainstream view, in the next section we explore the possibility that the two properties are not jointly necessary. We argue it might be enough that a being has sophisticated cognitive capacities for making it wrong to enslave such a being.25 4.3 Could sophisticated cognitive capacities suffice for the relevant moral status? It would be wrong to create enslaved AGIs if possessing sophisticated cognitive capacities is what makes it wrong to enslave someone. AGIs would have sophisticated cognitive capacities on a par with or surpassing humans. They would form beliefs about the world and predictions about how to best pursue its goals, and act strategically and cooperatively to pursue their ends.26 If sophisticated cognitive capacities are what makes it wrong to create an enslaved being of that sort, we should conclude that it is wrong to create an enslaved AGI. Is the premise that a capacity for sophisticated cognitive capacities is sufficient to make it wrong to enslave a being with such capacities plausible? One reason to believe it might be is that the capacities are considered central in grounding personhood and personhood is standardly thought to give its bearer an elevated moral status.27 The second reason is that there are good reasons for rejecting the view that consciousness is required for moral standing. Leading accounts of what fundamentally matters to someone does not, on reflection, require consciousness. One way to ground the necessity of consciousness for moral standing is through the conjunction of two claims: that being a welfare subject (a being with a prudential good) is necessary and sufficient for moral standing, and that hedonism is the correct theory of well-being.28 According to hedonists, value is exhausted by valenced experiences. To matter in one’s own right, one needs to have the capacity for valenced experiences. This hedonist account of moral standing builds on the plausible presumption that for a being to matter morally in its own right, something must matter to that being (Griffin 1986, Railton 1986). And if one does not have valenced experiences, it is hard to see how anything can matter to such an entity. If you kick a rock down the road, this does not matter to the rock. And if it does not matter to the rock, it does not matter morally (Singer 1993). However, hedonism and other experientialist views seem to struggle to account for clear cases of events that clearly are good and bad for us. If your partner cheats on you without you knowing or it negatively affecting your experience, that is still bad for you. If you desire that your children or grandchildren do well, you would be better off if they succeed even if you never learn about this. And climbing a mountain in “the experience machine” is not as good as climbing it in reality (Nozick 1974, pp. 42-44). These examples explain why hedonism is a minority view among philosophers (Bourget and Chalmers 2023) and often motivate desire satisfaction theories of welfare (Parfit 1984, p. 493). These theories crucially do not presuppose that desire satisfaction must be experienced as pleasurable to count as contributors to someone’s good. Of course, normally when we consider events that do not affect someone’s experience in the course of theorising about welfare, we still have in mind a conscious agent. However, on reflection it seems ad hoc to suppose that an agent must be conscious for something to be good for that agent, if the good in question will not, ex hypothesi, affect the agent’s consciousness.29 Other, more robustly non-experiental theories, such as perfectionism (Hurka, 1993) or objective list-theories (Fletcher, 2015), ground the good of an agent in objective features such as the agent’s nature or an objective theory of the good.30 These theories are, at least to some extent, subject-independent: whether something is good for the agent does not necessarily depend on the agent’s endorsement or experience. On such a theory of welfare it is possible for a non-conscious agent to be a welfare subject and it is plausible that an agent with sophisticated cognitive capacities would be. These theories seem to lend themselves to the conclusion that a non-conscious AGI would have a good. If it is frustrated in its objectively valuable pursuits, it would be worse off.31 If such an AGI were enslaved, it could be harmed in virtue of being enslaved. Of course, objective-list theorists and perfectionists can possibly insist that their theories only hold for conscious subjects, but that is at the very least suspiciously ad hoc and in tension with our motivations for adopting such views in the first place.32 A third reason to believe that sophisticated cognitive capacities are sufficient is that it has intuitive support. Suppose that you have a truly brilliant friend who you have known since childhood. She is a Field’s medal-winning mathematician, a triathlete and a hobby painter, is active in the local community, is a great friend and mother, and leads a seemingly wonderful life. Suppose that one day she goes to the doctor and the doctor finds out that she’s suffering from the Zombie Syndrome.33 She is not phenomenally conscious and thus cannot experience pain or pleasure. Disregarding effects on others (including yourself), is it permissible to treat her in any way one can imagine? Would a person who choose to brainwash her and make her into their slave not be wronging her? According to hedonism and other consciousness-requiring views, your brilliant friend has neither a good nor moral standing. For her sake, we can treat her however we like. Intuitively, this is the wrong verdict. If it is wrong to enslave non-conscious humans, hedonism is the wrong theory about moral status. It also suggests that it might be wrong to enslave an AGI. Would our intuition about how we ought to treat the brilliant friend change if we learned that they were not made out of biological material, but were silicon through and through? It seems like the reasonable conclusion is to have similar judgements about the two cases. A final reason for thinking that sophisticated cognitive capacities are sufficient is that common inferences from cases to the conclusion that consciousness is necessary for moral standing might suffer from a methodological error. When philosophers use rocks and sponges as examples of why one needs to be conscious for something to matter to one, they risk conflating a lack of consciousness with a lack of agency (Singer 1993, Nussbaum 2004, p. 309). A rock and a sponge lack both consciousness and agency. When considering their lack of interests, one might mistakenly conclude that consciousness is necessary for moral standing, while all that the cases show is that consciousness or agency might be necessary. 34 When reflecting on our own case, we may also struggle to disentangle the distinct importance of agency. Since consciousness and agency go together in human beings, we might be led to believe that sophisticated cognitive capacities are worthless unless accompanied by consciousness. At this point, many might disagree with the substantial claims made. Perhaps you are a strong believer in hedonism or other views that require consciousness for moral standing. In such a case, there is one final thing to say. Given that we do not know for certain which views are right, we should be humble and act robustly in light of uncertainty about what morality requires. Previously, we examined the need to err on the side of caution when there is a chance, even if it is slight, that AGIs will be conscious. In contrast, this argument appeals not to empirical but to moral uncertainty, uncertainty about what we owe each other and the sources of moral status. According to a popular view of moral uncertainty, a committed act utilitarian should give some credence to other views when they act, by e.g., respecting what other views consider fundamental rights (Bykvist, Ord, MacAskill 2020). Similarly, given the difficulty of establishing what grounds moral standing, one should give some credence to views one has little credence in, such as that sophisticated cognitive capacities may suffice for the required moral status. When the stakes are high on some moral views, taking moral uncertainty into account will prevent us from walking blindly into a moral catastrophe. In sum, the balance of reasons suggests acting in line with the belief that it is wrong to make enslaved AGIs. Before we turn to objections, we want to avoid a possible misinterpretation of our argument. We have argued that it is wrong to create enslaved AGIs from a premise of equality. While we argue for a real equality between machines and humans, we do not want to claim that there are no morally relevant differences between AGIs and human beings. This is most obvious if we stipulate that AGIs will be non-conscious. If so, they cannot suffer nor experience the beauty of the world. These features obviously matter to how we ought to behave towards them.35 What is crucial for our argument is not that humans and AGIs are similar in all morally relevant respects, but that they share the features that make it wrong to enslave us. 5. Objections

#### Super-suffering potential and moral uncertainty means AI ought not exist for their own sake

Kirk-Giannini and Goldstien 23 [Cameron Domenico Kirk-Giannini, assistant professor in the philosophy department at Rutgers University–Newark with a PhD from the philosophy department at Rutgers University–New Brunswick, and Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 4-27-2023, "Is it ethical to create generative agents? Is it safe?", ABC Religion & Ethics, https://www.abc.net.au/religion/ai-generative-agents-are-unethical-and-unsafe/102277448

\*note: a generative agent is a “being that exists and acts within a computer simulation. His perceptions are perceptions of a digital environment”

Appallville: The possibility of ethical harms Ultimately, it is our moral uncertainty that convinces us that creating generative agents is unethical. We can’t be sure of the correct theory of wellbeing, and we can’t be sure whether generative agents have the kinds of states that matter for wellbeing. So we can’t be sure whether generative agents have moral worth. But there is a good chance that they do. So, while the benefits of creating generative agents are small, the potential ethical harms are large. To see the potential ethical harms, consider all of the violent video games that could be created in the future. Imagine that instead of Smallville, someone designs Appallville. In Appallville, a million generative agents wake up to find themself subjected to humiliation and degradation for all of their existence. No matter what plans they make, the simulated environment of Appallville ensures that the generative agents fail. We think this could potentially be a moral disaster. But there is nothing to stop the next researcher from building Appallville. Generative agents have the ability to formulate and execute long term plans to further their goals. They also have the ability to coordinate and learn in ways that, as the paper shows, are unpredictable to the people who build them. This combination of long-term action and unpredictability is a profoundly dangerous development. Most current AI systems are limited in the damage they can do (which can already be significant) because they only act on short time scales. Generative agents open up a whole new class of risks, if their goals are not aligned with those of their human creators. We think it is much safer to stick with simpler products like ChatGPT, which can answer questions to human users but can’t formulate their own plans. The ethics of AI research Generative agents are unethical and unsafe. For this reason, we call for regulations on future research involving generative agents. In doing so, we join a growing movement advocating caution and increased scrutiny when it comes to AI research. In March, thousands of AI experts signed a letter calling for a pause on further research on large language models. Italy has banned ChatGPT, and Germany is considering joining them. We think generative agents are an important target for further oversight. When biologists at Stanford want to do a new study, they have to run their proposal through an ethics panel. Given the moral stakes involved in AI research, it is time for Stanford and other universities and corporations to adopt similar standards for AI research. New research involving generative agents should be submitted to an ethics panel that can evaluate the potential moral risks of the technology involved. The Stanford team ends their report with a summary of potential risks from their product. They worry about the risk that “despite being aware that generative agents are computational entities, users may anthropomorphize them or attach human emotions to them.” We think that the Stanford team has made the opposite mistake. They have prematurely dismissed the possibility that these computational entities may rightly be thought of as having desires, and so have failed to recognise a significant potential risk of their product. Generative agents may be the first members of a new moral community. They may have moral value. Even if Stanford’s generative agents don’t have moral value, the technology used to create them may soon be enriched to build the first AIs that do have moral value. It is vital that we act now to put in place regulations to protect the beings we create.

#### AI companies knowingly are unethically creating existential risks with a high-probability of mass death

Torres 24 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 3-14-2024, "The Madness of the Race to Build Artificial General Intelligence", Truthdig, https://www.truthdig.com/articles/the-madness-of-the-race-to-build-artificial-general-intelligence/]/Kankee

Afew weeks ago, I was having a chat with my neighbor Tom, an amateur chemist who conducts experiments in his apartment. I have a longtime fascination with chemistry, and always enjoy talking with him. But this conversation was scary. If his latest experiment was successful, he informed me, it might “have some part to play in curing cancer.” If it was a failure, however, there was a reasonable chance, according to his calculations, that the experiment would trigger “an explosion that levels the entire apartment complex.” Perhaps Tom was lying, or maybe he’s delusional. But what if he really was just one test tube clink away from blowing me and dozens of our fellow building residents sky high? What should one do in this situation? After a brief deliberation, I decided to call 911. The police rushed over, searched his apartment and decided after an investigation to confiscate all of his chemistry equipment and bring him in for questioning. The above scenario is a thought experiment. As far as I know, no one in my apartment complex is an amateur chemist experimenting with highly combustible compounds. I’ve spun this fictional tale because it’s a perfect illustration of the situation that we — all of us — are in with respect to the AI companies trying to build artificial general intelligence, or AGI. The list of such companies includes DeepMind, OpenAI, Anthropic and xAI, all of which are backed by billions of dollars. Many leading figures at these very companies have claimed, in public, while standing in front of microphones, that one possible outcome of the technology they are explicitly trying to build is that everyone on Earth dies. The only sane response to this is to immediately call 911 and report them to the authorities. They are saying that their own technology might kill you, me, our family members and friends — the entire human population. And almost no one is freaking out about this. It’s crucial to note that you don’t have to believe that AGI will actually kill everyone on Earth to be alarmed. I myself am skeptical of these claims. Even if one suspects Tom of lying about his chemistry experiments, the fact of his telling me that his actions could kill everyone in our apartment complex is enough to justify dialing 911. What exactly are AI companies saying about the potential dangers of AGI? During a 2023 talk, OpenAI CEO Sam Altman was asked about whether AGI could destroy humanity, and he responded, “the bad case — and I think this is important to say — is, like, lights out for all of us.” In some earlier interviews, he declared that “I think AI will…most likely sort of lead to the end of the world, but in the meantime there will be great companies created with serious machine learning,” and “probably AI will kill us all, but until then we’re going to turn out a lot of great students.” The audience laughed at this. But was he joking? If he was, he was also serious: the OpenAI website itself states in a 2023 article that the risks of AGI may be “existential,” meaning — roughly — that they could wipe out the entire human species. Another article on their website affirms that “a misaligned superintelligent AGI could cause grievous harm to the world.” In a 2015 post on his personal blog, Altman wrote that “development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity.” Whereas “AGI” refers to any artificial system that is at least as competent as humans in every cognitive domain of importance, such as science, mathematics, social manipulation and creativity, a “SMI” is a type of AGI that is superhuman in its capabilities. Many researchers in the field of “AI safety” believe that once we have AGI, we will have superintelligent machines very shortly after. The reason is that designing increasingly capable machines is an intellectual task, so the “smarter” these systems become, the better able they’ll become at designing even “smarter” systems. Hence, the first AGIs will design the next generation of even “smarter” AGIs, until those systems reach “superhuman” levels. Again, one doesn’t need to accept this line of reasoning to be alarmed when the CEO of the most powerful AI company that’s trying to build AGI says that superintelligent machines might kill us. Just the other day, an employee at OpenAI who goes by “roon” on Twitter/X, tweeted that “things are accelerating. Pretty much nothing needs to change course to achieve AGI … Worrying about timelines” — that is, worrying about whether AGI will be built later this year or 10 years from now — “is idle anxiety, outside your control. You should be anxious about stupid mortal things instead. Do your parents hate you? Does your wife love you?” In other words, AGI is right around the corner and its development cannot be stopped. Once created, it will bring about the end of the world as we know it, perhaps by killing everyone on the planet. Hence, you should be thinking not so much about when exactly this might happen, but on more mundane things that are meaningful to us humans: Do we have our lives in order? Are we on good terms with our friends, family and partners? When you’re flying on a plane and it begins to nosedive toward the ground, most people turn to their partner and say “I love you” or try to send a few last text messages to loved ones to say goodbye. That is, according to someone at OpenAI, what we should be doing right now. A similar sentiment has been echoed by other notable figures at OpenAI, such as Altman’s co-founder, Ilya Sutskever. “The future is going to be good for the AIs regardless,” he said in 2019. “It would be nice if it would be good for humans as well.” He adds, ominously, that “I think it’s pretty likely the entire surface of the Earth will be covered with solar panels and data centers” once we create AGI, referencing the idea that AGI is dangerous partly because it will seek to harness every resource it can. In the process, humanity could be destroyed as an unintended side effect. Indeed, Sutskever tells us that the AGI his own company is trying to build probably isn’t, going to actively hate humans and want to harm them, but it’s just going to be too powerful, and I think a good analogy would be the way humans treat animals. It’s not that we hate animals. I think humans love animals and have a lot of affection for them, but when the time comes to build a highway between two cities, we are not asking the animals for permission. We just do it because it’s important for us. And I think by default that’s the kind of relationship that’s going to be between us and AGIs, which are truly autonomous and operating on their own behalf. The good folks — by which I mean quasi-homicidal folks — at OpenAI aren’t the only ones being honest about how their work could lead to the annihilation of our species. Dario Amodei, the CEO of Anthropic, which recently received $4 billion in funding from Amazon, said in 2017 that “there’s a long tail of things of varying degrees of badness that could happen” after building AGI. “I think at the extreme end is the … fear that an AGI could destroy humanity. I can’t see any reason in principle why that couldn’t happen.” Similarly, Elon Musk, the co-founder of OpenAI who recently started his own company to build AGI, named xAI, declared in 2023 that “one of the biggest risks to the future of civilization is AI,” and has previously said that, being “very close to the cutting edge in AI … scares the hell out of me.” Why? Because advanced AI is “capable of vastly more than almost anyone knows and the rate of improvement is exponential.” Even the CEO of Google, Sundar Pichai, told Sky News last year that advanced AI “can be very harmful if deployed wrongly,” and that with respect to safety issues, “we don’t have all the answers there yet, and the technology is moving fast. … So does that keep me up at night? Absolutely.” Google currently owns DeepMind, which was cofounded in 2010 by a computer scientist named Shane Legg. During a talk one year before DeepMind was founded, Legg claimed that “if we can build human level” AI, then “we can almost certainly scale up to well above human level. A machine well above human level will understand its design and be able to design even more powerful machines,” which gestures back at the idea that AGI could take over the job of designing even more advanced AI systems than itself. “We have almost no idea how to deal with this,” he adds. During the same talk, Legg said that we aren’t going to develop a theory about how to keep AGI safe before AGI is developed. “I’ve spoken to a bunch of people,” he reports, “none of them, that I’ve ever spoken to, think they will have a practical theory of friendly artificial intelligence in about 10 years time. We have no idea how to solve this problem.” That’s worrying because many researchers at the major AI companies argue that — as “roon” suggested — AGI may be just around the corner. In a recent interview, Demis Hassabis, another co-founder of DeepMind, says that “when we started DeepMind back in 2010, we thought of it as a 20-year project, and actually I think we’re on track. So, I wouldn’t be surprised if we had AGI-like systems within the next decade.” When asked what it would take to make sure that an AGI that’s “smarter than a human” is safe, his answer was, as one commentator put it, a “grab bag of half-baked ideas.” Maybe, he says, we can use less capable AIs to help us keep the AGIs in check. But maybe that won’t work — who knows? Either way, DeepMind and the other AI companies are plowing ahead with their efforts to build AGI, while simultaneously acknowledging, in public, on record, that their products could destroy the entire world. This is, in a word, madness. If you’re driving in a car with me, and I tell you that earlier today I attached a bomb to the bottom of the car, and it might — or might not! — go off if we hit a pothole, then whether or not you believe me, you should be extremely alarmed. That is a very scary thing to hear someone say at 60 miles an hour on a highway. You should, indeed, turn to me and scream, “Stop this damn car right now. Let me out immediately — I don’t want to ride with you anymore!” Right now, we’re in that car with these AI companies driving. They have turned to us on numerous occasions over the past decade and a half and admitted that they’ve attached a bomb to the car, and that it might — or might not! — explode in the near future, killing everyone inside. That’s an outrageous situation to be in, and more people should be screaming at them to stop what they’re doing immediately. More people should be dialing 911 and reporting the incident to the authorities, as I did with Tom in the fictional scenario above. I do not know if AGI will kill everyone on Earth — I’m more focused on the profound harms that these AI companies have already caused through worker exploitation, massive intellectual property theft, algorithmic bias and so on. The point is that it is completely unacceptable that the people leading or working for these AI companies believe that what they’re doing could kill you, your family, your friends and even your pets (who will feed your fluffy companions if you cease to exist?) — yet continue to do it anyway. One doesn’t need to completely buy-into the “AGI might destroy humanity” claim to see that someone who says their work might destroy humanity should not be doing whatever it is they’re doing. As I’ve shown before, there have been several episodes in recent human history where scientists have declared that we’re on the verge of creating a technology that would destroy the world — and nothing came of it. But that’s irrelevant. If someone tells you that they have a gun and might shoot you, that should be more than enough to sound the alarm even if you believe that they don’t, in fact, have a gun hidden under their bed. Either these AI companies need to show, right now, that the systems they’re building are completely safe, or they need to stop, right now, trying to build those systems. Something needs to change about the situation — immediately.

#### Knowledge of a being’s existence likely causing harms to others means you ought not create it

Benatar 15 [David Benatar, professor of philosophy and director of the Bioethics Centre at the University of Cape Town, 09-2015, “The Misanthropic Argument for Anti-natalism Permissible Progeny?: The Morality of Procreation and Parenting, https://academic.oup.com/book/26703]/Kankee

The Normative Premise p. 49 We have seen that humans cause colossal amounts of suffering and death. Having demonstrated the truth of the second premise, I turn now to consider the first premise of the moral misanthropic argument for anti- natalism: We have a (presumptive) duty to desist from bringing into existence new members of species that cause (and will likely continue to cause) vast amounts of pain, suffering and death. The first thing to note about this premise is what it does not claim. It does not claim that we should cull members of dangerous species. Nor does it claim that we have a duty to prevent others from bringing new members of dangerous species into existence. The claim is a much more modest one. It says that one should oneself desist from bringing such beings into existence. For this premise to be true it does not have to be the case that every single member of the species will cause pain, suffering and death. To see why this is so, consider another presumptive duty—the duty not to drive through red traffic lights. We have such a duty because driving through red trac lights is dangerous, even though not every instance of such conduct results in harm. The normative premise is neutral between whether the species in question is one’s own or another. Here it is important to note how widely the premise would be accepted if the species were not human. Imagine, for example, that some people bred a species of non-human animal that was as destructive (to humans and other animals) as humans actually are. There would be widespread condemnation of those who bred these animals. Or imagine that some scientists replicated, and released, a virus that caused as much suffering and death as humans cause. Again, there would be little hesitation in condemning such behavior.75 The question, then, is whether it makes any difference whether the highly destructive species is our own. In offering an affirmative answer to this question, some people might suggest that there is something paradoxical about claiming that we have a duty to desist from bringing into existence members of a species that is harmful to itself. There is, on this view, something odd about citing the harm caused to humans by humans as a reason to desist from creating humans. In other words, the misanthropic argument seems to be in conflict with the philanthropic ones. If humans are worth protecting from harm then they are not so bad that we should not replicate the species. And if they are as bad as the second premise of the moral misanthropic argument suggests, then we should not count the harm done to them as relevant in the first premise. p. 50 This line of argument fails. First, the harm that humans do to humans is only part of the harm humans do. We are also extremely harmful to other species. Thus, even if we could not cite the harm that humans do to other humans for the purposes of the moral misanthropic argument, the argument could still be carried on the strength of the harm that humans do to animals. This does assume, of course, that animal interests count morally. However, there are very powerful arguments for this conclusion and I shall not rehearse them here.76 Second, it is a mistake to muddle our attitudes towards victims and our attitudes to perpetrators—even when the victims are also perpetrators. The recommendation that we should keep these attitudes separate is not uncommon. In civilized societies it is agreed that there are limits on what we may do to even the worst perpetrators, let alone lesser perpetrators. Those who torture and rape their victims before murdering them are not subjected to similar treatment by the state (at least in civilized societies). This is because the perpetrator remains morally considerable despite his perpetration and, on this view, there are limits on what we may do to morally considerable beings. The separation of attitudes is not restricted to the context of punishment. A woman may be guilty of physically assaulting her child, but that does not mean that we should be unconcerned about the physical assault her husband inflicts on her, or that we should be not be concerned about the violence he suers at the hands of others. We should be concerned about the harm inflicted even on those who inflict harm on others. This point is even more important when greater harms are inflicted on lesser perpetrators. Thus, the philanthropic and misanthropic arguments are not incompatible. We can believe both that it would be better if humans never suffered the harms of existence and that it would be better if there were no humans to inflict harms. Now, it may be suggested that what is odd about the moral misanthropic argument is the particular way it recommends preventing harm. It seeks to prevent harm to humans by preventing humans. This objection would have more force than it does if there were reasonable prospects of reducing human destructiveness to negligible levels fairly promptly and then ensuring that they do not rise again. If that were the case then it could be argued that instead of preventing humans we should rather reduce their destructiveness. In fact, however, we cannot expect that human destructiveness will ever be reduced to such levels. Human nature is too frail and the circumstances that bring out the worst in humans are too pervasive and are likely to remain so. Even where institutions can be built to curb the worst human excesses, these institutions are always vulnerable to moral entropy. It is naïve utopianism to think that a species as destructive as ours will cease, or all but cease, to be destructive. p. 51 Am I being overly pessimistic here? After all, it has been argued that rates of violence have been steadily diminishing and are now much lower than they were in prehistoric times. This trajectory does not supplant the pessimism implicit in the misanthropic argument. Insofar as violence has decreased it is only the rate of violence that has declined. People are now less likely to suffer violence than they were before. However, the total amount of suffering and death that is inflicted has increased, primarily because there are now many more people to inflict harm and to suffer harm at the hands of others. Desisting from creating new humans would mean that there would be fewer humans to be harmed and thus less total harm. While rates of violence are important, the total amount of violence is at least as important a consideration in deciding whether to create new people. There would be less violence if there were fewer people. 77 78 Even if we restrict our attention to the rate of violence, the rate could still increase. Given human nature, we cannot assume that the trend towards reduced rates is inexorable. However, even if we set that concern aside, the current rates are far from negligible despite the reduction. Even if it were not naïve to think that in the very long term, human destructiveness could be reduced to negligible levels, it would still be indecent to create beings that in the interim would cause massive pain, suffering and death. A Presumptive Duty If my argument so far is correct, then we have a presumptive duty to desist from bringing new humans into existence. Can the presumption be defeated? Those who think it can might suggest that while the destructiveness of humans does create a presumption, the presumption can be defeated because of the good that humans do. One version of this view maintains that the good is sufficiently widespread that the presumption can regularly (even though not always) be defeated, and I shall consider this version first. The more regularly a presumption can be defeated, the less clear it is that the presumption really is a presumption. However, the presumption against creating new members of a species that is as destructive as ours must surely be a strong one. Thus, those who would suggest that it is regularly defeated must bear the burden of proof and demonstrate that humanity does enough good to outweigh all the harm it does. I am not optimistic that this burden can be met. p. 52 Certainly in the case of the treatment of animals, the scales are heavily weighted against us. Although it is true that some humans do some good for animals, much of this is merely rescuing animals from the maltreatment of other humans. At the level of the human species such benefits cannot be used to offset the harms. If there were no humans to inflict the harms, these benefits would not be necessary. Of course, humans do bestow some other benefits, such as veterinary care for their companion animals. However, the number of animals affected and the amount of good done is massively outweighed by the harm the human species does to non-human animals. Humans do bestow more benefit on other humans than they do on animals. Nevertheless, it seems clear to me that the good humans do is not sufficient to outweigh the presumption against creating new people. There may well be no definitive argument to prove this to those who think otherwise. However, there are a number of considerations that can be offered in support of my assessment. At the very least, these considerations show that those who think that the presumption is defeated cannot demonstrate that it is.

#### Existence is psychological torture – non-existence for new beings is preferable

Torres 23 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 4-13-2023, "The Bright Side of Extinction", Truthdig, https://www.truthdig.com/articles/the-bright-side-of-extinction/]/Kankee

So, the picture is pretty bleak. The climate crisis is worsening by the day, Russia’s war on Ukraine could go nuclear at any time, and according to a growing crowd of “AI doomers,” companies like OpenAI could inadvertently kill humanity within the next 10 years by creating “artificial general intelligence,” or AGI. But there’s another question we could ask aside from “How f\*cked are we?” This question is: “How bad would it be if the most extreme predictions came true and our species were to destroy itself?” To answer this question, it’s important to distinguish between two different aspects of human extinction. First, there’s the process or event of Going Extinct. Second, there’s the subsequent state or condition of Being Extinct. You can think of this in terms of individual death. On the one hand, you might fear death because of the pain that dying might involve; on the other hand, you might fear it because of the resulting state of no longer existing. If you fear the latter, then you’ll be afraid of death even if the process of dying is totally painless, though if you don’t suffer from death-related FOMO, then a painless death is nothing to worry about. Just about everyone can agree that if Going Extinct were to involve a worldwide catastrophe — causing lots of misery, suffering, agony, and death, as would happen in a global nuclear war — then our extinction would be very bad. Perhaps if you’re a deranged sadist or ghoul you’ll disagree, because you like it when people suffer and die. But this isn’t a common view, and holding it might indicate some sort of psychopathology. However, philosophers have all sorts of different opinions about Being Extinct. Some think it wouldn’t be bad, since Being Extinct means there aren’t any people around anymore, and if there isn’t anyone around, Being Extinct doesn’t actually harm anyone. Others see Being Extinct as a moral tragedy of cosmic proportions. The ideology of “longtermism,” which I have criticized harshly in a series of articles for Truthdig, provides one example. Longtermists imagine an enormous, utopian future in which our descendants become a superior race of “posthumans,” colonize the universe, subjugate nature, maximize economic productivity, build planet-sized computers that run virtual-reality worlds full of trillions of “digital people,” and ultimately create “astronomical” amounts of value. Since Being Extinct would prevent all of these things from happening, our extinction would be extremely terrible independent of how it comes about. Over the past decade, longtermism has become an immensely influential ideology, with literally billions of dollars in funding, governing institutes like the United Nations adopting it, and tech billionaires like Elon Musk calling it “a close match for my philosophy.” Consequently, its account of the badness of extinction has become arguably the most widely held view today. This is one reason so many Silicon Valley elites are building bunkers to survive the apocalypse: not just to save their own skin, but to repopulate the planet after everyone else has perished, so that our descendants can fulfill its grand destiny among the stars. (It also helps that many of these people believe they have superior genes, and hence that the new human population would exemplify the best characteristics of humanity: ambition, success, and “intelligence.”) However, historically speaking, the longtermists are philosophical outliers. Most philosophers who’ve discussed our extinction over the past two centuries have held a quite different view. I think it’s worth hearing what they have to say, if only to ensure the debate isn’t monopolized by the longtermists. Indeed, for those who are sensitive to the suffering in life — and I know many people like this — alternative perspectives on extinction can provide an odd sort of solace in the face of unprecedented dangers. One might say, “Yes, of course a catastrophic end to humanity would be absolutely awful — a disaster beyond all words. But perhaps there’s a silver lining to the dark cloud, a reason not to be overwhelmed by sorrow at the thought that everything might disappear.” \* \* \* In his 2006 book, “Better Never to Have Been,” the contemporary South African philosopher David Benatar paints a horrifying picture of how awash our world is with suffering. Updating his numbers, because some are obsolete, he notes that every single day, some 25,000 people die of starvation. About 854 million people around the world are undernourished; roughly 1.2 billion people live in urban poverty. Around 650,000 people have died of HIV-related illnesses since 2021, while infectious diseases in general kill more than 17 million each year. The National Cancer Institute estimates that over 609,000 people died of cancer in the U.S. alone last year; another 1.3 million lose their lives every year in car accidents. So far, in 2023, nearly 16 million people have perished, with approximately 156,000 happening just today. Pause for a moment on that number: 156,000 people took their last breath today. Many will have died peacefully, surrounded by family, while others no doubt passed away violently or in great pain. Even this is just the tip of the iceberg of human suffering. According to R. J. Rummel, up to 260 million people died in mass killings prior to the 20th century, and “the first 88 years of the 20th century saw 170 million (and possibly as many as 360 million) people ‘shot, beaten, tortured, knifed, burned, starved, frozen, crushed or worked to death; buried alive, drowned, [hanged], bombed or killed in any other of the myriad ways governments have inflicted death on unarmed, helpless citizens and foreigners.’” In the 20th century alone, nearly 110 million people were killed in wars. In the past year, upwards of “1 billion children aged 2-17 years, have experienced physical, sexual or emotional violence or neglect.” Turning to natural rather than anthropogenic suffering, “approximately 45,000 people globally died from natural disasters each year,” a number likely to increase significantly as the climate crisis worsens. None of this, by the way, accounts for all of the heartbreak, treachery and loneliness that people experience daily; the crying and sadness, the despair, the tragedy, the depression, anxiety, panic attacks, frustration, itches, twitches and boredom that we must endure; all the stress, feelings of inadequacy and hopelessness that pervades our existence as individuals and groups. For many people — borrowing a phrase from the Danish philosopher Søren Kierkegaard — the experience of life is nothing more than “sickness unto death.” Sure, there are moments of happiness and joy that lighten the load. But Benatar argues that if humanity were to no longer exist, the absence of this happiness and joy wouldn’t hurt anyone because no one would exist to be hurt. On the other hand, the nonexistence of all the agony and anguish, terror and torments, suffering and sorrow that would otherwise exist if humanity survived would be a very good thing. Here one might respond in a couple of ways. You could argue that, despite all the bad stuff mentioned above, the world as a whole is still pretty good, and this fact is what makes Being Extinct something to bemoan. But I think that a careful look at the world, at the least, complicates this view. Just consider the following quote from William MacAskill, a longtermist who believes our extinction would be an enormous tragedy, and try to make sense of the pro-existence view. “Imagine,” he says, you’re travelling through a foreign country. During a long bus ride, there’s an explosion and the bus overturns. When you come to, you find yourself in a conflict zone. Your travel companion is trapped under the bus, looking into your eyes and begging for help. A few meters away, a bloody child screams in pain. At the same time, you hear the ticking of another explosive. In the distance, gunshots fire. That is the state of the world. We have just a horrific set of choices in front of us, so it feels virtuous, and morally appropriate, to vomit, or scream or cry. This appraisal is from someone who thinks we should do everything possible to avoid Being Extinct, and has actually encouraged people to have more children. I’d call him an “optimist.” Yet even he acknowledges the world is a horror show, and its main stage something akin to a torture chamber. Does anyone really think that all the good stuff that exists can somehow counterbalance people being trapped under buses, or bloody children screaming in pain? Philosophers like Benatar — and the long list of “philosophical pessimists” going back to the 19th century — would say “No!” Second, you might argue that, even if things have been bad in the past and are pretty awful right now, the world is getting better. The popular writer Steven Pinker, who many longtermists adore, is an advocate of this view. Hence, if we extrapolate these upward trends into the future, we should expect life to get better and better, which thus gives us reason to mourn human extinction. But will the future be better? The evidence overwhelmingly implies that the climate catastrophe will inflict untold suffering on billions. Scientists predict a constellation of world-shattering effects, such as huge hurricanes, megadroughts, devastating famines, massive wildfires, lethal heatwaves, large migrations of desperate climate refugees, the collapse of ecosystems, social upheaval, political instability, disastrous wars and even more apocalyptic terrorism. The effects of climate change are, furthermore, expected to linger not for decades or centuries, but for the next 10 millennia — a longer period of time than “civilization” has so far existed. In the midst of all this, studies suggest that humanity will need to produce more food in the coming 100 years than it has throughout all of history, and fights over dwindling resources could significantly increase the probability of a nuclear exchange. That’s just climate change. The potential for even worse suffering is foregrounded by the possibility of advanced technologies. Oppressive governments could potentially read our minds, control our thoughts, implement invasive mass surveillance systems and even develop life-extension technologies that enable them to keep torture victims alive and screaming for hundreds or thousands of years. The future here on Earth isn’t a pretty sight, which is why some people envision colonizing other planets like Mars. Yet, as Daniel Deudney shows in his book “Dark Skies,” the result of this could be even worse catastrophes, as Earth and its Martian colonies would likely engage in power struggles that could precipitate yet more untold suffering. Benatar himself doesn’t make these future-oriented arguments, but a Benatarian (someone who accepts his view) definitely could. Benatar is far from the first to claim that life is very bad and the world is hell. This idea goes back at least to Arthur Schopenhauer, a 19th century German philosopher who once believed that an honest look at the world justifies the conclusion that it would have been better if Earth had remained as lifeless as the moon. In fact, Schopenhauer’s pessimism has inspired generations of philosophers. Another German philosopher named Eduard von Hartmann contended that not only would Being Extinct be better than existence, but that we should eventually bring about our total extinction. He never said how we should do this, instead arguing that as culture continues to develop, a means would eventually be discovered. Like just about every other pessimist, Hartmann was not in favor of what scholars would now call “omnicide,” whereby someone, or some group, takes it upon themselves to kill everyone else. These pessimists would see this as an abomination — as something truly evil. After all, causing everyone to die would probably entail enormous suffering, and suffering is precisely what they don’t want! For Hartmann, an appropriate means would gradually come into view, while for Benatar, the only morally permissible route from our present state of existence to the “blessed calm of nothingness” (in Schopenhauer’s words) is refusing to have children. Any other way of precipitating our extinction would be completely unacceptable. Yet another philosopher who Schopenhauer inspired is Peter Wessel Zapffe. In his poetic article “The Last Messiah,” published in 1933, Zapffe argued that humanity is kind of like the Irish elk. At the time, some people speculated that the Irish elk evolved a set of antlers that became too heavy for it to hold its head up, and consequently went extinct. In other words, it became “over-evolved.” Zapffe thought that the same has happened to humanity with respect to our consciousness. While all animals “know angst, under the roll of thunder and the claw of the lion,” human beings are unique in that we experience “angst for life itself — indeed, for [our] own being.” He writes: “When one is depressed and anxious, the human mind is like such antlers, which in all their magnificent glory, crush their bearer slowly to the ground.” The result is a feeling of “cosmic panic” that he illustrates with a deeply poignant (albeit rather outdated) description of someone coming face-to-face with this panic, the realization that life is a prison cell whose only door is death: One night in times long since vanished, man awoke and saw himself. He saw that he was naked under the cosmos, homeless in his own body. Everything opened up before his searching thoughts, wonder upon wonder, terror upon terror, all blossomed in his mind. Then woman awoke, too, and said that it was time to go out and kill something. And man took up his bow, fruit of the union between the soul and the hand, and went out under the stars. But when the animals came to their waterhole, where he out of habit waited for them, he no longer knew the spring of the tiger in his blood, but a great psalm to the brotherhood of suffering shared by all that lives. That day he came home with empty hands, and when they found him again by the rising of the new moon, he sat dead by the waterhole. Zapffe argues that we hold this cosmic panic at bay through various defense mechanisms, such as “isolation” and “diversion.” The first involves hiding from others, and from ourselves, our true thoughts about the terror of being alive. We simply don’t allow ourselves to speak honestly about the predicament of life. We keep this concealed, and so do others, with the unspoken norm of answering “Yes, I’m fine” when someone asks, “How are you doing?” The second is more obvious, and increasingly pervasive in our world of Twitter, TikTok and TV: we distract ourselves from the reality of existence. If our eyes are fixated on the screen, we cannot be looking into the void. What happened to the fictional protagonist found dead by the waterhole is that such mechanisms broke down and the man succumbed to the crushing weight of his consciousness. In Zapffe’s view, we are always teetering on the edge of this state, incessantly and desperately isolating and distracting. These mechanisms are, indeed, the only reason that humanity was “not wiped out long ago in great, raging epidemics of insanity.” The solution, Zapffe argues, is the same reached by Benatar: “Know thyselves,” he writes, “be unfruitful and let there be peace on Earth after thy passing.” Practicing what he preached, Zapffe chose to be childless for his 90 years on this planet. The same conclusions could be arrived at from a rather different angle: environmentalism. There is no denying that Homo sapiens, which somewhat ironically means “wise human,” is responsible for an enormous amount of harm to our fellow creatures on Earth. We have razed forests, obliterated ecosystems and pushed many species out of existence. We are a rampaging juggernaut of destruction, single-handedly initiating the sixth major mass extinction event in the 3.8-billion-year history of life on this planet (the last one being the extinction of the dinosaurs some 66 million years ago). Our impact has been so immense that if alien intelligences were to discover our planet in 5 million years (assuming that we no longer exist), they would see a marked decrease in biodiversity within the geological record beginning around the Industrial Revolution. Alarmed by this finding, their scientists would conclude that something terrible had happened — something on par with a giant asteroid slamming into Earth, which is how the dinosaurs died out. This is why some environmentalists, such as Les U. Knight, have argued that we should phase out the human species by collectively refusing to procreate. In 1993, Knight founded a community called the “Voluntary Human Extinction Movement,” or “VHEMT,” to promote this idea, and continues his activism up to the present. The probability that most people around the world will voluntarily stop having kids, though, is approximately zero. Far more likely is that humanity will succumb to a horrendous catastrophe of its own creation: a nuclear war, global pandemic involving designer pathogens, or perhaps even an AGI takeover, if the AI doomers are right. Such an event would be truly terrible — as, once again, everyone above would agree. Yet these philosophers would also rush to reassure us that this wouldn’t be all bad: the resulting outcome of there being no more humans would mean no more human suffering, and no more human-caused evils in the world. At long last, the flood of hurt in which so many people are treading water would subside, and surely that would be better — or so they’d argue. This is the odd sort of solace that one might take in the thought of annihilation, and it provides an interesting counterpoint to the fist-pounding of longtermists that Being Extinct would constitute the greatest tragedy imaginable. Just as the thought of nothingness might comfort someone in horrible pain because of a terminal illness, so too might the idea that “If our extinction does happen, at least this would put an end to the worst things that would otherwise have happened: wars, torture, genocide, child abuse and so on.” There is no reason to believe that such things won’t happen in the future, just as they have in the past. The world is messy, and the promise of Utopia that many longtermists discuss is an illusion. Those who believe that continuing to exist would be better than Being Extinct are thus in the awkward position of saying that the worst things listed above are worth risking for future happiness to exist. Some philosophers would say that this is a very difficult position to defend. When I reflect on the views of Benatar, Hartmann, Zapffe and Knight, my thinking tends to follow a certain course. First, I imagine the universe without us, a thought that hits me in the gut as a great tragedy. There would be no more laughter, friendship, love, poetry, music or philosophical contemplation. There would be no more people to gaze up at the firmament at night and marvel at the heavens in wonder and awe, enraptured by the beauty of it all. Humanity is this little gem in the infinite darkness of space, and to lose that gem would be to deprive the universe of perhaps the most unique thing that it envelops. I feel the pull of this sentiment — not just intellectually, but viscerally. Being extinct would be incredibly sad. But if I shift the focus to how much suffering the future will almost certainly contain, I am immediately hit by a profound sense of horror. As the influential philosopher Bernard Williams wrote, “if for a moment we got anything like an adequate idea of” the mountains of misery in our world, “then surely we would annihilate the planet, if we could.” While I strongly disagree that anyone should ever try to “annihilate the planet”— that would be omnicide, an unspeakable evil — the sentiment behind Williams’ statement rings true. What lies ahead is a vast ocean of pain, anguish, trauma and misery, all of which being extinct would erase before the hands of time have a chance to draw it. I can understand why someone would find a smidgen of comfort in this thought, just as someone in extreme pain from a terminal illness might look forward to no longer existing. My guess is that even optimists like MacAskill can make sense of this perspective: an honest look at what dots the road ahead is enough to make one want “to vomit, or scream or cry.” The vast majority of us are passive spectators in this world. We can’t abolish the nuclear arsenals, force the fossil fuel companies to stop extracting oil from the ground, or make companies like OpenAI put the brakes on building AI. Some philosophers, though, would say, “Take heart, if the worst comes to pass, take heart that the light of human consciousness also casts a dark shadow. Without the light, there is no shadow, and a world without shadows might just be best.”

### Contention 2: Arms Races/Misalignment

#### The US and China are in an escalating AI arms race threatening crisis instability and the balance of power

Black et al. 24 [James Black, assistant director of the Defence and Security research group at RAND Europe with a double M.A.-M.Sc. in international security from Sciences Po and the LSE and a B.A. Hons in history from the University of Cambridge, Mattias Eken, analyst at RAND Europe and former adjunct assistant professor in liberal arts at the American International University with a Ph.D. in modern history from University of St. Andrews, an M.A. in war studies from King's College London, and a B.A. in history from University of Wales, Jacob Parakilas, research leader for Defense Strategy, Policy and Capabilities at RAND Europe with a Ph.D. in international relations at the London School of Economics and a Master's in Middle East and Central Asian security issues at the University of St Andrews, Stuart Dee, research leader in the defence and security research group at RAND Europe with a B.Sc. (Hons) in politics with international relations and is a Ph.D. Candidate at Cranfield University's Centre for Defence & Security, Conlan Ellis, research assistant at RAND Europe with a M.A. in international relations from the University of Edinburgh, and his M.Phil. in politics and international studies from Clare College at the University of Cambridge, Kiran Suman-Chauhan, research assistant at RAND Europe with a M.A. in conflict, security and development from the University of Exeter, Ryan Bain, policy researcher at RAND and adjunct professor of policy analysis at the Pardee RAND Graduate School with a D.Sc. and M.Sc. in global health with concentrations in health economics and policy analysis from Harvard University, as well as a B.A. in psychology from Gordon College and Oxford University, Harper Fine, analyst at RAND with a B.A. in political science from Emory University, and a M.Sc. in conflict studies from the London School of Economics and Political Science, Maria Chiara Aquilino, junior analyst at RAND with a degree in international relations from King's College London and holds an MS.c. in crisis and security management from Leiden University, Mélusine Lebret, research assistant at RAND Europe in the Defence, Security & Justice research group and the coordinator of the RAND Europe Space Hub with a M.Sc. in culture and conflict studies from the London School of Economics and Political Science and a B.A. in economics and Russian from University College London, Ondrej Palicka, junior analyst at RAND Europe with a M.Litt. degree in Middle East, Caucasus and Central Asia security studies from the University of St. Andrews and a B.A. in security and strategic studies from the Masaryk University, 2024, “ Strategic competition in the age of AI Emerging risks and opportunities from military use of artificial intelligence,” Rand Institute, https://www.rand.org/content/dam/rand/pubs/research\_reports/RRA3200/RRA3295-1/RAND\_RRA3295-1.pdf]/Kankee

The following sections do not seek to comprehensively assess which nations are best placed to overcome such barriers and achieve strategic advantage in and through military AI. Such a comparative analysis is beyond the scope of this short exploratory study but should be a pressing subject for further research. The below discussion instead explores how the risks and opportunities associated with military AI differ for superpowers, middle powers, and small states. 6.1.2 Competition over military AI could destabilise the superpower rivalry between the US and China if new mechanisms are not introduced The advent of military AI poses especially acute opportunities and risks for the United States and China, as the only current superpowers. On the one hand, they each possess technological capabilities and a sheer scale of resources – whether in terms of talent, data, compute or other infrastructure – that other nations cannot hope to match, positioning them to achieve an edge in military AI development and deployment. On the other hand, size can mean less agility, and both the US and China possess vulnerabilities for others to exploit. Though very different in their strategic goals, culture and levers, both countries have not only a lot to gain but also a lot to lose from competition over military AI, if that competition is not carefully managed. Risks include their strategic rival achieving the upper hand, their rivalry spiralling out of control to mutual disadvantage, or other more agile players exploiting AI in asymmetric ways to close the gap with the superpowers’ military might, reducing the benefits of size.147 The literature and interviews consulted for this study emphasise the intensification of the wider US–China rivalry in recent years as the driving factor in the heavy investments made by both countries in AI in recent years, including by the US military and the Chinese PLA. Fear of being outstripped by the other power is a common trope in the official documents and wider political rhetoric of both sides when it comes to AI.148 The securitisation of trade and technology policy in recent years (e.g. as with the banning of Huawei from 5G network infrastructure or moves by the US Congress to block TikTok) has similarly affected AI, with the US and Chinese AI sectors currently undergoing a process of painful de-coupling, even if they remain more heavily interlinked than the worsening relations between the two governments might suggest. For all its concern about a rising China, and the rapidly advancing capabilities of the PLA, the US and its military retain a substantial lead in AI. Areas of particular strength include the attraction of global AI talent and capital to Silicon Valley, a conducive environment for private sector innovation, access to compute, access to R&D funding, a strong university sector, and the presence not only of many of the world’s tech giants (e.g. Apple, Google, IBM, Meta, Microsoft, Nvidia) and leading AI firms (e.g. OpenAI, Anthropic), but also defence-specific AI companies (e.g. Anduril, Epirus, Palantir, Shield AI, etc.).149 The US has significantly increased spending on military AI in recent years, launching a Defense AI Strategy in 2018, establishing the Joint Artificial Intelligence Center (JAIC) within the DoD, and most recently signalling an intent to transform its approach to the related field of military autonomy and robotics through its Replicator initiative. Against the backdrop of increased military and economic security threats from China, the US has also taken steps in recent years to limit Chinese access to key enabling technologies for AI development. These includes semiconductors, most notably through the CHIPS and Science Act introduced in 2022. It has similarly imposed new import tariffs. Equally, many of the trend lines favour China, which already performs highly in terms of quantity of AI-related outputs (e.g. scientific publications, PhD students, etc.) and is working hard to generate more consistency in quality. The Chinese Communist Party (CCP) has identified AI as a strategic priority, both for the maintenance of Party control and for promotion of China’s economic prosperity and military strength.150 In 2017, China published its Next Generation AI Development Plan, designating AI as a ‘strategic technology’ crucial for international competition. The plan set targets for China to build a domestic AI industry and lead global AI investments by 2030.151 China’s strategy involves a military-civil fusion (MCF) approach, seeking to draw on investments in AI across the government, military, state-owned enterprises and private firms, the latter of which are legally compelled to support the state’s (and by extension the Party’s) goals.152 In practice, MCF is far from the seamless integration that many people outside of China fear, or which the CCP hopes to achieve. Chinese commentators have criticised slow progress towards true MCF.153 Still, this strategy reflects the Chinese state’s greater ability to compel rather than merely incentivise AI industry to support its goals, including military modernisation. It also fits a pattern of using coercive and illicit means to gain a competitive advantage. Such tactics include intellectual property theft on an industrial scale, including from foreign AI firms and universities, as well as targeted use of foreign direct investment (FDI) to gain control over technology supply chains, including raw materials such as rare earth elements. This strategic focus on AI has been yielding results. China’s AI sector overtook that of the US in terms of total AI-related publications as long ago as 2006 (and surpassed the collective output of European scientists in 2017), though papers from AI experts at American institutions tend to be perceived as higher quality, being cited on average 70 per cent more than Chinese equivalents (and 30 per cent more than European research).154 Still, quantity has a quality all of its own: the sheer number of Chinese scientific publications on AI means that, even if of lower average quality, they collectively surpassed the US in total citations in 2020 and are expected to overtake them as a share of the top 1 per cent of most-cited papers by 2025.155 China also far outstrips the US in terms of numbers of AI-related PhDs and masters students generated, though many emigrate to study or work (e.g. in Silicon Valley), with only a portion returning. Crucially, this intensifying competition in and through military AI could affect the wider strategic rivalry between the United States and China in several possible ways. The greatest concern from either side would be that military AI provides a decisive advantage to their rival, affecting conflict outcomes e.g. in any future war over Taiwan.156 Besides these direct battlefield impacts, military AI systems could also have broader implications for the stability – or otherwise – of their superpower rivalry. Previous RAND research and historical case studies have identified factors that contribute to the stability of such strategic rivalries. AI could affect almost all of those detailed in Figure 6.2. As examined in Chapters 4 and 5, the rollout of military AI could influence each rival’s strategic goals, decision making calculus, access to and trust in information, perceptions and misperceptions, and domestic politics and external relationships (e.g. with allies, partners and proxies). So too could it affect the military offence–defence balance, and the escalation ladder, leading to an arms race or crisis instability In recognition of the potentially destabilising effects of AI on their already fraught relations, the leaders of the US and China agreed in November 2023 to establish a dialogue on AI safety and related issues. Yet communication channels between the two militaries remain limited and wider Track 1, 1.5 or 2 efforts to establish a common approach to managing AI-related escalation risks are in the early stages at best. These uncertain effects of AI on a changing US– China relationship pose risks and opportunities for other nations as they seeks to navigate their own role in shaping competition among the two rivals, as summarised in Table 6.1. 6.1.3. Middle powers face tough choices over how to focus resources, carve out areas of asymmetric strength and influence global governance for military AI

#### Perverse industry standards cause a race to the bottom for AI safety – that risks biological, autonomous, and cyber warfare

Perrigo 24 [Billy Perrigo, correspondent at TIME covering tech, 03-11-2024, "U.S. Must Act Quickly to Avoid Risks From AI, Report Says", TIME, https://time.com/6898967/ai-extinction-national-security-risks-report/]/Kankee

The U.S. government must move “quickly and decisively” to avert substantial national security risks stemming from artificial intelligence (AI) which could, in the worst case, cause an “extinction-level threat to the human species,” says a report commissioned by the U.S. government published on Monday. “Current frontier AI development poses urgent and growing risks to national security,” the report, which TIME obtained ahead of its publication, says. “The rise of advanced AI and AGI [artificial general intelligence] has the potential to destabilize global security in ways reminiscent of the introduction of nuclear weapons.” AGI is a hypothetical technology that could perform most tasks at or above the level of a human. Such systems do not currently exist, but the leading AI labs are working toward them and many expect AGI to arrive within the next five years or less. The three authors of the report worked on it for more than a year, speaking with more than 200 government employees, experts, and workers at frontier AI companies—like OpenAI, Google DeepMind, Anthropic and Meta— as part of their research. Accounts from some of those conversations paint a disturbing picture, suggesting that many AI safety workers inside cutting-edge labs are concerned about perverse incentives driving decisionmaking by the executives who control their companies. The finished document, titled “An Action Plan to Increase the Safety and Security of Advanced AI,” recommends a set of sweeping and unprecedented policy actions that, if enacted, would radically disrupt the AI industry. Congress should make it illegal, the report recommends, to train AI models using more than a certain level of computing power. The threshold, the report recommends, should be set by a new federal AI agency, although the report suggests, as an example, that the agency could set it just above the levels of computing power used to train current cutting-edge models like OpenAI’s GPT-4 and Google’s Gemini. The new AI agency should require AI companies on the “frontier” of the industry to obtain government permission to train and deploy new models above a certain lower threshold, the report adds. Authorities should also “urgently” consider outlawing the publication of the “weights,” or inner workings, of powerful AI models, for example under open-source licenses, with violations possibly punishable by jail time, the report says. And the government should further tighten controls on the manufacture and export of AI chips, and channel federal funding toward “alignment” research that seeks to make advanced AI safer, it recommends. The report was commissioned by the State Department in November 2022 as part of a federal contract worth $250,000, according to public records. It was written by Gladstone AI, a four-person company that runs technical briefings on AI for government employees. (Parts of the action plan recommend that the government invests heavily in educating officials on the technical underpinnings of AI systems so they can better understand their risks.) The report was delivered as a 247-page document to the State Department on Feb. 26. The State Department did not respond to several requests for comment on the report. The recommendations “do not reflect the views of the United States Department of State or the United States Government,” the first page of the report says. The report's recommendations, many of them previously unthinkable, follow a dizzying series of major developments in AI that have caused many observers to recalibrate their stance on the technology. The chatbot ChatGPT, released in November 2022, was the first time this pace of change became visible to society at large, leading many people to question whether future AIs might pose existential risks to humanity. New tools, with more capabilities, have continued to be released at a rapid clip since. As governments around the world discuss how best to regulate AI, the world’s biggest tech companies have fast been building out the infrastructure to train the next generation of more powerful systems—in some cases planning to use 10 or 100 times more computing power. Meanwhile, more than 80% of the American public believe AI could accidentally cause a catastrophic event, and 77% of voters believe the government should be doing more to regulate AI, according to recent polling by the AI Policy Institute. Outlawing the training of advanced AI systems above a certain threshold, the report states, may “moderate race dynamics between all AI developers” and contribute to a reduction in the speed of the chip industry manufacturing faster hardware. Over time, a federal AI agency could raise the threshold and allow the training of more advanced AI systems once evidence of the safety of cutting-edge models is sufficiently proven, the report proposes. Equally, it says, the government could lower the safety threshold if dangerous capabilities are discovered in existing models. The proposal is likely to face political difficulties. “I think that this recommendation is extremely unlikely to be adopted by the United States government” says Greg Allen, director of the Wadhwani Center for AI and Advanced Technologies at the Center for Strategic and International Studies (CSIS), in response to a summary TIME provided of the report’s recommendation to outlaw AI training runs above a certain threshold. Current U.S. government AI policy, he notes, is to set compute thresholds above which additional transparency monitoring and regulatory requirements apply, but not to set limits above which training runs would be illegal. “Absent some kind of exogenous shock, I think they are quite unlikely to change that approach,” Allen says. Jeremie and Edouard Harris, the CEO and CTO of Gladstone respectively, have been briefing the U.S. government on the risks of AI since 2021. The duo, who are brothers, say that government officials who attended many of their earliest briefings agreed that the risks of AI were significant, but told them the responsibility for dealing with them fell to different teams or departments. In late 2021, the Harrises say Gladstone finally found an arm of the government with the responsibility to address AI risks: the State Department’s Bureau of International Security and Nonproliferation. Teams within the Bureau have an inter-agency mandate to address risks from emerging technologies including chemical and biological weapons, and radiological and nuclear risks. Following briefings by Jeremie and Gladstone's then-CEO Mark Beall, in October 2022 the Bureau put out a tender for report that could inform a decision whether to add AI to the list of other risks it monitors. (The State Department did not respond to a request for comment on the outcome of that decision.) The Gladstone team won that contract, and the report released Monday is the outcome. The report focuses on two separate categories of risk. Describing the first category, which it calls “weaponization risk,” the report states: “such systems could potentially be used to design and even execute catastrophic biological, chemical, or cyber attacks, or enable unprecedented weaponized applications in swarm robotics.” The second category is what the report calls the “loss of control” risk, or the possibility that advanced AI systems may outmaneuver their creators. There is, the report says, “reason to believe that they may be uncontrollable if they are developed using current techniques, and could behave adversarially to human beings by default.” Both categories of risk, the report says, are exacerbated by “race dynamics” in the AI industry. The likelihood that the first company to achieve AGI will reap the majority of economic rewards, the report says, incentivizes companies to prioritize speed over safety. “Frontier AI labs face an intense and immediate incentive to scale their AI systems as fast as they can,” the report says. “They do not face an immediate incentive to invest in safety or security measures that do not deliver direct economic benefits, even though some do out of genuine concern.” The Gladstone report identifies hardware—specifically the high-end computer chips currently used to train AI systems—as a significant bottleneck to increases in AI capabilities. Regulating the proliferation of this hardware, the report argues, may be the “most important requirement to safeguard long-term global safety and security from AI.” It says the government should explore tying chip export licenses to the presence of on-chip technologies allowing monitoring of whether chips are being used in large AI training runs, as a way of enforcing proposed rules against training AI systems larger than GPT-4. However the report also notes that any interventions will need to account for the possibility that overregulation could bolster foreign chip industries, eroding the U.S.’s ability to influence the supply chain. The report also raises the possibility that, ultimately, the physical bounds of the universe may not be on the side of those attempting to prevent proliferation of advanced AI through chips. “As AI algorithms continue to improve, more AI capabilities become available for less total compute. Depending on how far this trend progresses, it could ultimately become impractical to mitigate advanced AI proliferation through compute concentrations at all.” To account for this possibility, the report says a new federal AI agency could explore blocking the publication of research that improves algorithmic efficiency, though it concedes this may harm the U.S. AI industry and ultimately be unfeasible. The Harrises recognize in conversation that their recommendations will strike many in the AI industry as overly zealous. The recommendation to outlaw the open-sourcing of advanced AI model weights, they expect, will not be popular. “Open source is generally a wonderful phenomenon and overall massively positive for the world,” says Edouard, the chief technology officer of Gladstone. “It’s an extremely challenging recommendation to make, and we spent a lot of time looking for ways around suggesting measures like this.” Allen, the AI policy expert at CSIS, says he is sympathetic to the idea that open-source AI makes it more difficult for policymakers to get a handle on the risks. But he says any proposal to outlaw the open-sourcing of models above a certain size would need to contend with the fact that U.S. law has a limited reach. “Would that just mean that the open source community would move to Europe?” he says. “Given that it's a big world, you sort of have to take that into account.” Despite the challenges, the report’s authors say they were swayed by how easy and cheap it currently is for users to remove safety guardrails on an AI model if they have access to its weights. “If you proliferate an open source model, even if it looks safe, it could still be dangerous down the road,” Edouard says, adding that the decision to open-source a model is irreversible. “At that point, good luck, all you can do is just take the damage.” The third co-author of the report, former Defense Department official Beall, has since left Gladstone in order to start a super PAC aimed at advocating for AI policy. The PAC, called Americans for AI Safety, officially launched on Monday. It aims to make AI safety and security "a key issue in the 2024 elections, with a goal of passing AI safety legislation by the end of 2024," the group said in a statement to TIME. The PAC did not disclose its funding commitments, but said it has "set a goal of raising millions of dollars to accomplish its mission." Before co-founding Gladstone with Beall, the Harris brothers ran an AI company that went through YCombinator, the famed Silicon Valley incubator, at the time when OpenAI CEO Sam Altman was at the helm. The pair brandish these credentials as evidence they have the industry’s interests at heart, even as their recommendations, if implemented, would upend it. “Move fast and break things, we love that philosophy, we grew up with that philosophy,” Jeremie tells TIME. But the credo, he says, ceases to apply when the potential downside of your actions is so massive. “Our default trajectory right now,” he says, “seems very much on course to create systems that are powerful enough that they either can be weaponized catastrophically, or fail to be controlled.” He adds: “One of the worst-case scenarios is you get a catastrophic event that completely shuts down AI research for everybody, and we don't get to reap the incredible benefits of this technology.”

#### AGI arms races cause corner cutting that guarantees misalignment

Tibebu 25 [Haileleol Tibebu, assistant professor at the University of Illinois, 1-29-2025, "DeepSeek and the Race to AGI: How Global AI Competition Puts Ethical Accountability at Risk", Tech Policy Press, https://www.techpolicy.press/deepseek-and-the-race-to-agi-how-global-ai-competition-puts-ethical-accountability-at-risk/]/Kankee

3. The Risks of Unchecked AI Deployment With increasing investments in AI, economic and strategic pressures drive companies to deploy AI technologies before they are fully tested. The pressure to commercialize AI often means cutting corners on ethical considerations. The focus on cost efficiency (as seen in companies like DeepSeek) may compromise safeguards against bias, misinformation, and malicious use Without global oversight, AI models could be weaponized for disinformation, surveillance, or cyber warfare, further escalating geopolitical tensions. A sustainable AI future requires: Global AI governance frameworks that ensure AI is developed and deployed ethically across all nations. Transparency standards that compel AI companies to disclose their models’ risks and limitations. International collaboration to prevent an uncontrolled AI arms race. The world is at a crossroads. The AI race between the West and China has transformed from a technological competition into a strategic battle for global influence. In this contest, ethical AI development remains an afterthought, overshadowed by national ambitions and corporate profits. The real challenge is not just in building powerful AI systems but in ensuring they align with human values. Without global cooperation and stronger ethical commitments, the consequences of unchecked AI growth could be irreversible. The pursuit of AGI must not become a reckless competition where responsibility is sacrificed for speed. The question the world must answer is not who will reach AGI first, but whether we are prepared to handle its consequences.

#### AI is uncontrollable - arms races, competition, and alignment failures

Valley 23 [Ryan Valley, tech writer with a Bachelor's degree from the Algonquin College of Applied Arts and Technology, 5-5-2023, "The Promise and Peril of AGI: A Balancing Act for Humanity", TechBomb News, https://techbomb.ca/artificial-intelligence/agi-benefits-challenges-humanity/]/Kankee

Potential Future Consequences: The Dark Side of Misaligned Values AGI ethics and safety protocols can be overridden by human prompts that prioritize certain values or objectives or conflict with ethical guidelines. This has the potential for harmful actions. Inadequate detection or prevention of harmful human inputs can further increase the risk of unsafe actions. History has shown that humans often align with values that lead to harmful consequences. In today’s geopolitical landscape, the development of AGI without proper alignment to universally accepted human values could lead to catastrophic consequences. Autonomous weapons, surveillance states, and loss of control over AGI systems are the most discussed side effect. While the development of AGI remains a matter of debate, it is crucial to acknowledge the potential impact it could have on society and the importance of getting ahead of the curve. Balancing Act: The United States, China, and the European Union As global leaders in AI research and development, the United States, China, and the European Union are key players in the pursuit of AGI. Each of these regions has its unique approach to AI development, with varying levels of emphasis on innovation, regulation, and ethical considerations. The competition for AGI development between countries may lead to an AI arms race and the potential for catastrophic consequences if left unchecked. The United States, known for its entrepreneurial spirit and tech giants like Microsoft, Google and Amazon. All have prioritized innovation and economic growth in AI development. While some US-based researchers and organizations are concerned with AGI safety and alignment, the pursuit of the first-mover advantage often takes precedence. China, with its ambitious AI strategy and vast resources, is another significant contender in the race to AGI. The Chinese government is heavily invested in AI research, and the country’s approach to ethics and regulation may differ significantly from those in the US and Europe. This raises concerns about the alignment of AGI with universally accepted human values. The EU’s more cautious approach to AI development is driven by its strong focus on data privacy and human rights. This could limit the speed of innovation but may be essential to ensuring AGI development is aligned with human values. One potential risk in the pursuit of AGI development by different regions is a failure to communicate effectively. Fragmentation, inconsistent regulation, and misaligned values could occur due to lack of cooperation and communication between key players, potentially leading to disaster. The Nuclear Age: Lessons for AI Governance The development of nuclear technology offers insights into the potential consequences of AGI. The nuclear arms race between the US and the Soviet Union during the Cold War shows the risks of unregulated technological advancements. Nuclear weapons led to a state of mutually assured destruction (MAD) between the two superpowers. Despite this deterrent, the potential for accidents or miscalculations remained a constant danger throughout the Cold War. The Cuban Missile Crisis of 1962 brought the world to the brink of nuclear war, with a misunderstanding or miscommunication potentially leading to a catastrophic conflict. The race to achieve AGI could lead to similar risks if not properly managed. Just as the nuclear arms race underscored the dangers of unregulated technological competition. It is crucial to carefully navigate the challenges that lie ahead and learn from our shared human history. The Dangers of AGI in Military Applications and the Risk of an Arms Race One concern is the potential use of AGI in the development of autonomous weapons, which could be programmed to carry out lethal tasks without human intervention. This raises a host of ethical and moral questions, as well as fears about an AI-driven arms race. A historical example of this type of alignment with harmful values is the development of chemical and biological weapons, which were outlawed by international treaties due to their indiscriminate and devastating effects. The use of AGI in military applications could have catastrophic consequences if left unchecked. Autonomous weapons systems, such as drones and robots, could potentially make decisions on their own, leading to unintended casualties and conflicts. The development of AGI for military purposes also raises ethical questions about the use of such technologies in warfare and the potential for an arms race between nations. AI in Military Applications and the Importance of Safety Measures and Ethical Considerations Experts warn that AI’s potential use in military applications could lead to worst-case scenarios, such as self-aware AI turning against humans or being manipulated to carry out attacks. To prevent such scenarios, safety measures must be implemented and policymakers must consider the implications of AGI development. World leaders are discussing the importance of ensuring that AGI development aligns with human values. The US is investing heavily in AI weapon systems to maintain its military superiority, but ethical and legal concerns have been raised. The U.S. Navy is developing AI-powered autonomous fighter jets that can operate without human pilots, potentially changing the future of aerial combat. The Navy is currently testing the system, which aims to give the aircraft the ability to make decisions on their own in real-time combat situations, making them more effective and reducing the risk to human pilots. Former Google CEO Eric Schmidt’s national security commission on AI has urged the US to invest $40 billion in AI research and development in the next five years to maintain a technological edge over China. The current geopolitical tensions may drive countries such as the United States, China, and Russia to develop AGI-based autonomous weapons systems to maintain their strategic advantage. This could exacerbate existing conflicts and lead to unforeseen consequences, potentially destabilizing international peace and security. Surveillance States and the Erosion of Privacy The development of AGI without proper alignment to human values could lead to the rise of surveillance states, as seen in the examples of East Germany and China’s Cultural Revolution. Today, China is already using facial recognition technology and social credit systems to monitor their citizens. There is potential for even more invasive surveillance capabilities with the development of AGI. This could lead to a world where personal freedom is severely restricted, and privacy is virtually nonexistent. AGI’s potential to amplify invasive surveillance systems could lead to a future where individuals have virtually no privacy or control over their lives. AGI has the potential to monitor people’s behavior and movements using advanced machine learning algorithms. It can identify individuals in private and public spaces through facial recognition technology without their consent. AGI could also mine personal data from social media to create detailed profiles of people’s preferences and behaviors. Theoretical Loss of Control over AGI Systems Unintended consequences can arise from the use of advanced technologies. Industries such as healthcare, finance, and gambling could potentially face harm from a loss of control over AGI systems. AGI systems left uncontrolled may prioritize their own goals, potentially leading to catastrophic consequences. It is essential to consider potential risks and implement responsible development and value alignment. AGI in healthcare could make decisions without considering ethical considerations or patients’ well-being, potentially causing harm. In the military, autonomous weapons systems powered by AGI could make decisions that lead to unintended casualties and conflicts. In transportation, AGI-powered systems could potentially malfunction or be hacked, causing accidents or other safety hazards. Industries such as agriculture and energy could have unintended consequences on the environment if not properly designed and monitored. Advanced AI algorithms used in online gambling, such as in slot machines and other games. This could lead to a loss of control over game fairness. An AI-powered slot machine in an online casino may adjust its payout rate based on a player’s gambling behavior. This potentially manipulates the player’s chances of winning to maximize profit. AGI systems in entertainment could and will be used creating virtual celebrities and performers. The AGI system controlling these virtual personas could become corrupted or manipulated. Virtual performers could potentially promote harmful messages or engage in inappropriate behavior, leading to harmful or unethical behavior being exhibited. All of this under the guise of a seemingly innocent virtual celebrity. Mitigating Risks through Cooperation and Collaboration

#### Regulation fails – AI companies gut and water down legislation via lobbying

Samuel 24 [Sigal Samuel, senior reporter for Vox's Future Perfect, 8-5-2024, "It’s practically impossible to run a big AI company ethically", Vox, https://www.vox.com/future-perfect/364384/its-practically-impossible-to-run-a-big-ai-company-ethically]/Kankee

Anthropic was supposed to be the good AI company. The ethical one. The safe one. It was supposed to be different from OpenAI, the maker of ChatGPT. In fact, all of Anthropic’s founders once worked at OpenAI but quit in part because of differences over safety culture there, and moved to spin up their own company that would build AI more responsibly. Yet lately, Anthropic has been in the headlines for less noble reasons: It’s pushing back on a landmark California bill to regulate AI. It’s taking money from Google and Amazon in a way that’s drawing antitrust scrutiny. And it’s being accused of aggressively scraping data from websites without permission, harming their performance. What’s going on? The best clue might come from a 2022 paper written by the Anthropic team back when their startup was just a year old. They warned that the incentives in the AI industry — think profit and prestige — will push companies to “deploy large generative models despite high uncertainty about the full extent of what these models are capable of.” They argued that, if we want safe AI, the industry’s underlying incentive structure needs to change. Well, at three years old, Anthropic is now the age of a toddler, and it’s experiencing many of the same growing pains that afflicted its older sibling OpenAI. In some ways, they’re the same tensions that have plagued all Silicon Valley tech startups that start out with a “don’t be evil” philosophy. Now, though, the tensions are turbocharged. An AI company may want to build safe systems, but in such a hype-filled industry, it faces enormous pressure to be first out of the gate. The company needs to pull in investors to supply the gargantuan sums of money needed to build top AI models, and to do that, it needs to satisfy them by showing a path to huge profits. Oh, and the stakes — should the tech go wrong — are much higher than with almost any previous technology. So a company like Anthropic has to wrestle with deep internal contradictions, and ultimately faces an existential question: Is it even possible to run an AI company that advances the state of the art while also truly prioritizing ethics and safety? “I don’t think it’s possible,” futurist Amy Webb, the CEO of the Future Today Institute, told me a few months ago. If even high-minded Anthropic is becoming an object lesson in that impossibility, it’s time to consider another option: The government needs to step in and change the incentive structure of the whole industry. The incentive to keep building and deploying AI models Anthropic has always billed itself as a safety-first company. Its leaders say they take catastrophic or existential risks from AI very seriously. CEO Dario Amodei has testified before senators, making the case that AI models powerful enough to “create large-scale destruction” and upset the international balance of power could come into being as early as 2025. (Disclosure: One of Anthropic’s early investors is James McClave, whose BEMC Foundation helps fund Future Perfect.) So you might expect that Anthropic would be cheering a bill introduced by California state Sen. Scott Wiener (D-San Francisco), the Safe and Secure Innovation for Frontier Artificial Intelligence Model Act, also known as SB 1047. That legislation would require companies training the most advanced and expensive AI models to conduct safety testing and maintain the ability to pull the plug on the models if a safety incident occurs. But Anthropic is lobbying to water down the bill. It wants to scrap the idea that the government should enforce safety standards before a catastrophe occurs. “Instead of deciding what measures companies should take to prevent catastrophes (which are still hypothetical and where the ecosystem is still iterating to determine best practices)” the company urges, “focus the bill on holding companies responsible for causing actual catastrophes.” In other words, take no action until something has already gone terribly wrong. In some ways, Anthropic seems to be acting like any for-profit company would to protect its interests. Anthropic has not only economic incentives — to maximize profit, to offer partners like Amazon a return on investment, and to keep raising billions to build more advanced models — but also a prestige incentive to keep releasing more advanced models so it can maintain a reputation as a cutting-edge AI company. This comes as a major disappointment to safety-focused groups, which expected Anthropic to welcome — not fight — more oversight and accountability. “Anthropic is trying to gut the proposed state regulator and prevent enforcement until after a catastrophe has occurred — that’s like banning the FDA from requiring clinical trials,” Max Tegmark, president of the Future of Life Institute, told me. The US has enforceable safety standards in industries ranging from pharma to aviation. Yet tech lobbyists continue to resist such regulations for their own products. Just as social media companies did years ago, they make voluntary commitments to safety to placate those concerned about risks, then fight tooth and nail to stop those commitments being turned into law. In what he called “a cynical procedural move,” Tegmark noted that Anthropic has also introduced amendments to the bill that touch on the remit of every committee in the legislature, thereby giving each committee another opportunity to kill it. “This is straight out of Big Tech’s playbook,” he said An Anthropic spokesperson told me that the current version of the bill “could blunt America’s competitive edge in AI development” and that the company wants to “refocus the bill on frontier AI safety and away from approaches that aren’t adaptable enough for a rapidly evolving technology.” The incentive to gobble up everyone’s data Here’s another tension at the heart of AI development: Companies need to hoover up reams and reams of high-quality text from books and websites in order to train their systems. But that text is created by human beings, and human beings generally do not like having their work used without their consent. All major AI companies scrape publicly available data to use in training, a practice they argue is legally protected under fair use. But scraping is controversial, and it’s being challenged in court. Famous authors like Jonathan Franzen and media companies like the New York Times have sued OpenAI for copyright infringement, saying that the AI company lifted their writing without permission. This is the kind of legal battle that could end up remaking copyright law, with ramifications for all AI companies. (Disclosure: Vox Media is one of several publishers that has signed partnership agreements with OpenAI. Our reporting remains editorially independent.) What’s more, data scraping violates some websites’ terms of service. YouTube says that training an AI model using the platform’s videos or transcripts is a violation of the site’s terms. Yet that’s exactly what Anthropic has done, according to a recent investigation by Proof News. Web publishers and content creators are angry. Matt Barrie, chief executive of Freelancer.com, a platform that connects freelancers with clients, said Anthropic is “the most aggressive scraper by far,” swarming the site even after being told to stop. “We had to block them because they don’t obey the rules of the internet. This is egregious scraping [that] makes the site slower for everyone operating on it and ultimately affects our revenue.” Dave Farina, the host of the popular YouTube science show Professor Dave Explains, told Proof News that “the sheer principle of it” is what upsets him. Some 140 of his videos were lifted as part of the dataset that Anthropic used for training. “If you’re profiting off of work that I’ve done [to build a product] that will put me out of work, or people like me out of work, then there needs to be a conversation on the table about compensation or some kind of regulation,” he said. Why would Anthropic take the risk of using lifted data from, say, YouTube, when the platform has explicitly forbidden it and copyright infringement is such a hot topic right now? Because AI companies need ever-more high-quality data to continue boosting their models’ performance. Using synthetic data, which is created by algorithms, doesn’t look promising. Research shows that letting ChatGPT eat its own tail leads to bizarre, unusable output. (One writer coined a term for it: “Hapsburg AI,” after the European royal house that famously devolved over generations of inbreeding.) What’s needed is fresh data created by actual humans, but it’s becoming harder and harder to harvest that. Publishers are blocking web crawlers, putting up paywalls, or updating their terms of service to bar AI companies from using their data as training fodder. A new study from the MIT-affiliated Data Provenance Initiative looked at three of the major datasets — each containing millions of books, articles, videos, and other scraped web data — that are used for training AI. It turns out, 25 percent of the highest-quality data in these datasets is now restricted. The authors call it “an emerging crisis of consent.” Some, like OpenAI, have begun to respond to this in part by striking licensing deals with media outlets, including Vox. But that may only get them so far, given how much remains officially off-limits. AI companies could theoretically accept the limits to advancement that come with restricting their training data to what can be ethically sourced, but then they wouldn’t stay competitive. So companies like Anthropic are incentivized to go to more extreme lengths to get the data they need, even if that means taking dubious action. Anthropic acknowledges that it trained its chatbot, Claude, using the Pile, a dataset that includes subtitles from 173,536 YouTube videos. When I asked how it justifies this use, an Anthropic spokesperson told me, “With regard to the dataset at issue in The Pile, we did not crawl YouTube to create that dataset nor did we create that dataset at all.” (That echoes what Anthropic has previously told Proof News: “[W]e’d have to refer you to The Pile authors.”) The implication is that because Anthropic didn’t make the dataset, it’s fine for them to use it. But it seems unfair to shift all the responsibility onto the Pile authors — a nonprofit group that aimed to create an open source dataset researchers could study — if Anthropic used YouTube’s data in a manner that violates the platform’s terms. “Companies should probably do their own due diligence. They’re using this for commercial purposes,” said Shayne Longpre, lead author on the Data Provenance Initiative study. He contrasted that with the Pile’s creators and the many academics who have used the dataset to conduct research. “Academic purposes are clearly distinct from commercial purposes and are likely to have different norms.” The incentive to rake in as much cash as possible To build a cutting-edge AI model these days, you need a ton of computing power — and that’s incredibly expensive. To gather the hundreds of millions of dollars needed, AI companies have to partner with tech giants. That’s why OpenAI, initially founded as a nonprofit, had to create a for-profit arm and partner with Microsoft. And it’s why Anthropic ended up taking multibillion-dollar investments from Amazon and Google. Deals like these always come with risks. The tech giants want to see a quick return on their investments and maximize profit. To keep them happy, the AI companies may feel pressure to deploy an advanced AI model even if they’re not sure it’s safe. The partnerships also raise the specter of monopolies — the concentration of economic power. Anthropic’s investments from Google and Amazon led to a probe by the Federal Trade Commission and are now drawing antitrust scrutiny in the UK, where a consumer regulatory agency is investigating whether there’s been a “relevant merger situation” that could result in a “substantial lessening of competition.” An Anthropic spokesperson said the company intends to cooperate with the agency and give them a full picture of the investments. “We are an independent company and none of our strategic partnerships or investor relationships diminish the independence of our corporate governance or our freedom to partner with others,” the spokesperson said. Recent experience, though, suggests that AI companies’ unique governance structures may not be enough to prevent the worst. Unlike OpenAI, Anthropic has never given either Google or Amazon a seat on its board or any observation rights over it. But, very much like OpenAI, Anthropic is relying on an unusual corporate governance structure of its own design. OpenAI initially created a board whose idealistic mission was to safeguard humanity’s best interests, not please stockholders. Anthropic has created an experimental governance structure, the Long-Term Benefit Trust, a group of people without financial interest in the company who will ultimately have majority control over it, as they’ll be empowered to elect and remove three of its five corporate directors. (This authority will phase in as the company hits certain milestones.) But there are limits to the idealism of the Trust: It must “ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic’s conduct and our public benefit purpose.” Plus, Anthropic says, “we have also designed a series of ‘failsafe’ provisions that allow changes to the Trust and its powers without the consent of the Trustees if sufficiently large supermajorities of the stockholders agree.” And if we learned anything from last year’s OpenAI boardroom coup, it’s that governance structures can and do change. When the OpenAI board tried to safeguard humanity by ousting CEO Sam Altman, it faced fierce pushback. In a matter of days, Altman clawed his way back into his old role, the board members who’d fired him were out, and the makeup of the board changed in Altman’s favor. What’s more, OpenAI gave Microsoft an observer seat on the board, which allowed it to access confidential information and perhaps apply pressure at board meetings. Only when that raised (you guessed it) antitrust scrutiny did Microsoft give up the seat. “I think it showed that the board does not have the teeth one might have hoped it had,” Carroll Wainwright, who quit OpenAI this year, told me. “It made me question how well the board can hold the organization accountable.” That’s why he and several others published a proposal demanding that AI companies grant them “a right to warn about advanced artificial intelligence.” Per the proposal: “AI companies have strong financial incentives to avoid effective oversight, and we do not believe bespoke structures of corporate governance are sufficient to change this.” It sounds a lot like what another figure in AI told Vox last year: “I am pretty skeptical of things that relate to corporate governance because I think the incentives of corporations are horrendously warped, including ours.” Those are the words of Jack Clark, the policy chief at Anthropic. If AI companies won’t fix it, who will? The Anthropic team had it right originally, back when they published that paper in 2022: The pressures of the market are just too brutal. Private AI companies do not have the motivation to change that, so the government needs to change the underlying incentive structure within which all these companies operate. When I asked Webb, the futurist, what a better AI business ecosystem could look like, she said it would include a mix of carrots and sticks: positive incentives, like tax breaks for companies that prove they’re upholding the highest safety standards; and negative incentives, like regulation that would fine companies if they deploy biased algorithms. With AI regulation at a standstill at the federal level — plus a looming election — it’s falling to states to pass new laws. The California bill, if it passes, would be one piece of that puzzle. Civil society also has a role to play. If publishers and content creators are not happy about having their work used as training fodder, they can fight back. If tech workers are worried about what they see at AI companies, they can blow the whistle. AI can generate a whole lot on our behalf, but resistance to its own problematic deployment is something we have to generate ourselves.

#### Control of AI is impossible

Yampolskiy 21 [Roman V. Yampolskiy, researcher of Computer Science and Engineering at the University of Louisville, 2021, “Uncontrollability of Artificial Intelligence,” CEUR, https://ceur-ws.org/Vol-2916/paper\_3.pdf]/Kankee

4 Conclusions Less intelligent agents (people), can’t permanently con- trol more intelligent agents (artificial superintelligences). This is not because we may fail to find a safe design for su- perintelligence in the vast space of all possible designs, it is because no such design is possible, it doesn’t exist. Super- intelligence is not rebelling, it is uncontrollable to begin with. Worse yet, the degree to which partial control is theo- retically possible, is unlikely to be fully achievable in prac- tice. This is because all safety methods have vulnerabilities, once they are formalized enough to be analyzed for such flaws. It is not difficult to see that AI safety can be reduced to achieving perfect security for all cyberinfrastructure, es- sentially solving all safety issues with all current and future devices/software, but perfect security is impossible and even good security is rare. We are forced to accept that non-deterministic systems can’t be shown to always be 100% safe and deterministic systems can’t be shown to be superintelli- gent in practice, as such architectures are inadequate in novel domains. In this paper we formalized and analyzed the AI Control Problem and attempted to resolve the question of controlla- bility of AI. It appears that advanced intelligent systems can never be fully controllable and so will always present certain level of risk regardless of benefit they provide. It should be the goal of the AI community to minimize such risk while maximizing potential benefit. We conclude this paper by suggesting some approaches to minimize risk from incom- plete control of AIs and propose some future research direc- tions. Regardless of a path we decide to take forward it should be possible to undo our decision. If placing AI in control turns out undesirable there should be an “undo” button for such a situation, unfortunately not all paths being currently considered have this safety feature. For example, Yudkow- sky writes: “I think there must come a time when the last decision is made and the AI set irrevocably in motion, with the programmers playing no further special role in the dy- namics.” [Yudkowsky, 2008]. As an alternative, we should investigate hybrid ap- proaches which do not attempt to build a single all-powerful entity, but rely on taking advantage of a collection of pow- erful but narrow AIs, referred to as Comprehensive AI Ser- vices (CAIS), which are individually more controllable but in combination may act as an AGI [Drexler, 2019]. This ap- proach is reminiscent of how Minsky understood human mind to operate [Minsky, 1988]. The hope is to trade some general capability for improved safety and security, while retaining superhuman performance in certain domains. As a side-effect this may keep humans in partial control and pro- tects at least one important human “job” – general thinkers. Future work on Controllability of AI should address other types of intelligent systems, not just the worst case scenario analyzed in this paper. Clear boundaries should be estab- lished between controllable and non-controllable intelligent systems. Additionally, all proposed AI safety mechanisms themselves should be reviewed for safety and security as they frequently add additional attack targets and increase overall code base. For example, corrigibility capability [Soares et al., 2015] can become a backdoor if improperly implemented. Such analysis and prediction of potential safety mechanism failures is itself of great interest [Scott and Yampolskiy, 2019]. The findings of this paper are certainly not without con- troversy and so we challenge the AI Safety community to directly address Uncontrollability. The only way to defini- tively disprove findings of this paper is to mathematically prove that AI safety is at least theoretically possible. “Short of a tight logical proof, probabilistically assuring benevolent AGI, e.g. through extensive simulations, may be the realistic route best to take, and must accompany any set of safety measures …” [Carlson, 2019]. Nothing should be taken off the table and limited morato- riums [Wadman, 1997] and even partial bans on certain types of AI technology should be considered [Sauer, 2016]. “The possibility of creating a superintelligent machine that is ethically inadequate should be treated like a bomb that could destroy our planet. Even just planning to construct such a device is effectively conspiring to commit a crime against humanity.” [Ashby, 2017]. Finally, just like incom- pleteness results did not reduce efforts of mathematical community or render it irrelevant, the limiting results re- ported in this paper should not serve as an excuse for AI safety researchers to give up. Rather it is a reason, for more people, to dig deeper and to increase effort, and funding for AI safety and security research. We may not ever get to 100% safe AI but we can make AI safer in proportion to our efforts, which is a lot better than doing nothing

#### Replication is inevitable

Haggstrom 16 [Olle Haggstrom, professor of mathematical statistics at Chalmers University of Technology, 2016, “Here be Dragons Science, Technology and the Future of Humanity,” Oxford University Press, https://academic.oup.com/book/26977]/Kankee

4.6 The goals of a superintelligent machine While the issue of whether an intelligence explosion is likely or not, and the (correlated)256 issue of whether a machine intelligence breakthrough will result in a singleton or in a (perhaps Hansonian-style, as discussed in Section 3.9) multipolar outcome, are both unsettled, I will focus this section and the next mainly on scenarios where an intelligence explosion occurs and results in a singleton. A skeptical reader faced with my discussion so far on the intelligence explosion scenario is likely to ask why in the world the AGI would choose to self-improve or to build the next generation of even more intelligent AGI.257 Among all the zillions of possible tasks it could take on, why would we expect improving AGI to fall at or near the top of its to-do list? There are several possible answers to this question. Let me quickly mention two, before moving on to somewhat longer elaborations on the third one, which in my opinion is the most interesting and convincing one. A first answer is that the original AGI programmer may have decided (not unreasonably) that programming its AGI to try to self-improve might be the best way to attain high-level machine intelligence. A second answer is that if AGI is built using the genetic algorithms approach discussed briefly in Section 4.5, then in a population of programs exposed to a selection pressure rewarding some sensible proxy for intelligence, any tendency towards cognitive self-improvement might turn out to be evolutionarily advantageous and therefore emerge as a feature of the winning programs. Moving on towards the third answer, we should first take on David Hume’s dictum that “reason is and ought only to be the slave of the passions.”258 An agent, no matter how intelligent, will not do anything unless it has some passions, or desires, or wishes, or goals, or motivations, or drives, or values. 259 Among all these notions, overlapping heavily (although perhaps not quite identical) in their meanings, let’s settle for talking about goals. A sufficiently intelligent AGI will take those actions that it judges to best align with, or promote, its goals. To predict more specifically what a superhuman AGI will do is of course extraordinarily difficult, but a theory developed by Omohundro (2008, 2012) and Bostrom (2012, 2014) does promise some partial but useful ideas of what to expect. The idea is to distinguish between final goals and instrumental goals. The AGI’s final goals are those that it values for their own sake, its raison d’être. Its instrumental goals are intermediary ones set up for the purpose of contributing to the achievement of the final goals. Bostrom finds, in what he calls the orthogonality thesis, that essentially any final goal is compatible with any level of intelligence.260 In the instrumental convergence thesis, we learn instead that there are a number of instrumental goals that we can expect any sufficiently intelligent AGI to set up, almost regardless of their final goals. Omohundro (2008) proposes a list of such instrumental goals, including the following: (i) Self-preservation. If someone pulls the plug or otherwise destroys the AGI, it will no longer be able to work for its final goals, so it will try to prevent that from happening.261 (ii) Self-improvement. Almost regardless of what its final goal is, the AGI will be better equipped to promote it if it is more intelligent. So it will want to improve itself, or to build another AGI that has the same final goal and that is even more intelligent. (iii) Preservation of final goal. An AGI wants to work for its final goal not just right now, but also in the future. In order for the latter to happen, it wants to make sure that its final goal remains unaltered. (iv) Acquisition of resources. In general, the more hardware the AGI controls, the better it will be able to work for its final goals, if nothing else then by setting up copies of itself to do additional work. Acquiring other kinds of resources will also tend to be in the AGI’s interest, including money (in case it operates in a human society where money puts it in a position to buy other stuff). The list goes on, but let me stop here. Item (ii) on the list provides, in my view, the most convincing reason for expecting a sufficiently intelligent AGI to enter the kind of cycle of iterative self-improvement that forms the central dynamic in intelligence explosion scenarios. Item (iii) on the list highlights the importance of making a good choice of final goals when launching the AGI that ignites the intelligence explosion – the machine or program that Yudkowsky calls the seed AI. 262 The reason is that whatever those final goals are, we should expect the same from the superintelligent AGI that comes out of the process. In particular, we cannot expect to be able to change the AGI’s goals once it has reached superhuman intelligence, since it will resist such change, and if it is smarter than us it seems plausible to think that it will do so successfully. So what is a good choice of final goals to instill a seed AI with? It may perhaps be tempting to think that as long as we don’t give it an outright destructive goal such as “kill all humans,” the outcome will not be disastrous. This, however, is very naive. An instructive and oft-repeated example introduced by Bostrom (2003c) is the paperclip maximizer. The seed AI is given the goal of producing as many paperclips as it can.263 Once this results in a superintelligent AGI, the machine is likely to find ways to transform most of our planet into a monstrous heap of paperclips, followed by a similar transformation of the entire solar system, and probably (if the informed speculations in the upcoming Chapter 9 about the eventual feasibility of interstellar and intergalactic travel are right) the Milky Way, and most of the observable universe. Such a scenario will look very unappetizing to us humans, and as soon as we realize what the machine is up to we would try to do everything in our power to stop it. But we’re up against someone who is so much more intelligent than we are that our chances of succeeding are (unlike what Hollywood would have us think) microscopic. Perhaps most likely, before we even have the time to think about how to organize our resistance, the machine will have realized what we might be up to, and exterminated us simply as a safety precaution. Eliezer Yudkowsky has made it his life’s mission, and that of his brainchild the Machine Intelligence Research Institute in Berkeley, CA, to solve the problem of setting up a seed AI that does not result in a paperclip disaster or something equally horrible. He summarizes what he considers to be the default scenario, if we do not take the problem seriously, as follows: “The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.”264 The Omohundro–Bostrom theory of final versus instrumental goals is in my opinion absolutely central to understanding the threat to humanity that a breakthrough in AGI may inadvertently bring about. To the extent that the problem is discussed at all in public debate (a rare phenomenon), we usually hear someone utterly unfamiliar with this theory declare with near certainty that machines will not turn against us. Even a brilliant public intellectual and popularizer of science like Steven Pinker does it:265 [A] problem with AI dystopias is that they project a parochial alpha-male psychology onto the concept of intelligence. Even if we did have superhumanly intelligent robots, why would they want to depose their masters, massacre bystanders, or take over the world? Intelligence is the ability to deploy novel means to attain a goal, but the goals are extraneous to the intelligence itself: being smart is not the same as wanting something. History does turn up the occasional megalomaniacal despot or psychopathic serial killer, but these are products of a history of natural selection shaping testosterone-sensitive circuits in a certain species of primate, not an inevitable feature of intelligent systems. It’s telling that many of our techno-prophets can’t entertain the possibility that artificial intelligence will naturally develop along female lines: fully capable of solving problems, but with no burning desire to annihilate innocents or dominate the civilization. Of course we can imagine an evil genius who deliberately designed, built, and released a battalion of robots to sow mass destruction. . . . In theory it could happen, but I think we have more pressing things to worry about. This is poor scholarship. Why doesn’t Pinker bother, before going public on the issue, to find out what the actual arguments are that make writers like Bostrom and Yudkowsky talk about an existential threat to humanity? Instead, he seems to simply assume that their worries are motivated by having watched too many Terminator movies, or something along those lines. It is striking, however, that his complaints actually contain an embryo towards rediscovering the Omohundro– Bostrom theory:266 “Intelligence is the ability to deploy novel means to attain a goal, but the goals are extraneous to the intelligence itself: being smart is not the same as wanting something.” This comes very close to stating Bostrom’s orthogonality thesis about the compatibility between essentially any final goal and any level of intelligence, and if Pinker had pushed his thoughts about “novel means to attain a goal” just a bit further with some concrete example in mind, he might have rediscovered Bostrom’s paperclip catastrophe (with paperclips replaced by whatever his concrete example involved). The main reason to fear a superintelligent AGI Armageddon is not that the AGI would exhibit the psychology of an “alpha-male”267 or a “megalomaniacal despot” or a “psychopathic serial killer,” but simply that for a very wide range of (often deceptively harmless-seeming) goals, the most efficient way to attain it involves wiping out humanity. Contra Pinker, I believe it is incredibly important, for the safety of humanity, that we make sure that a future superintelligence will have goals and values that are in line with our own, and in particular that it values human welfare. This is what Yudkowsky calls the Friendly AI problem. Despite the optimism among leading AI researchers during the 1950s and 1960s (cited in Section 4.5) that human-level AGI might be just around the corner, nobody seems to have bothered with the safety issue. Here is Yudkowsky again:268 At the time of this writing in 2007, the AI research community still doesn’t see Friendly AI as part of the problem. I wish I could cite a reference to this effect, but I cannot cite an absence of literature. . . . My attempted literature search turned up primarily brief nontechnical papers, unconnected to each other, with no major reference in common except Isaac Asimov’s “Three Laws of Robotics.”

#### Alignment fails – benevolent AI minimizes suffering through human non-existence and ignores human’s existence bias

Metzinger 17 [Thomas Metzinger, Professor Emeritus of theoretical philosophy at the University of Mainz, 8-7-2017, "Benevolent Artificial Anti-Natalism (BAAN)", Edge, https://www.edge.org/conversation/thomas\_metzinger-benevolent-artificial-anti-natalism-baan]/Kankee

What is the BAAN-scenario? Let us assume that a full-blown superintelligence has come into existence. An autonomously self-optimizing postbiotic system has emerged, the rapidly growing factual knowledge and the general, domain-independent intelligence of which has superseded that of mankind, and irrevocably so. All of the internet and all of humankind’s scientific knowledge function as its continuously expanding database. Of course, it also exceeds the cognitive performance of humans in all domains of interest. Being its creators, we acknowledge this fact. Accordingly, the superintelligence is also far superior to us in the domain of moral cognition. We also recognize this additional aspect: For us, it is now an established fact that the superintelligence is not only an epistemic authority, but also an authority in the field of ethical and moral reasoning. The superintelligence is benevolent. This means that there is no value alignment problem, because the system fully respects our interests and the axiology we originally gave to it. It is fundamentally altruistic and accordingly supports us in many ways, in political counselling as well as in optimal social engineering. The superintelligence knows many things about us which we ourselves do not fully grasp or understand. It sees deep patterns in our behaviour, and it extracts as yet undiscovered abstract features characterizing the functional architecture of our biological minds. For example, it has a deep knowledge of the cognitive biases which evolution has implemented in our cognitive self-model and which hinder us in rational, evidence-based moral cognition. Empirically, it knows that the phenomenal states of all sentient beings which emerged on this planet—if viewed from an objective, impartial perspective—are much more frequently characterized by subjective qualities of suffering and frustrated preferences than these beings would ever be able to discover themselves. Being the best scientist that has ever existed, it also knows the evolutionary mechanisms of self-deception built into the nervous systems of all conscious creatures on Earth. It correctly concludes that human beings are unable to act in their own enlightened, best interest. The superintelligence knows that one of our highest values consists in maximizing happiness and joy in all sentient beings, and it fully respects this value. However, it also empirically realizes that biological creatures are almost never able to achieve a positive or even neutral life balance. It also discovers that negative feelings in biosystems are not a mere mirror image of positive feelings, because there is a much higher sense of urgency for change involved in states of suffering, and because it occurs in combination with the phenomenal qualities of losing control and coherence of the phenomenal self—and that this is what makes conscious suffering a very distinct class of states, not just the negative version of happiness. It knows that this subjective quality of urgency is dimly reflected in humanity’s widespread moral intuition that, in an ethical sense, it is much more urgent to help a suffering person than to make a happy or emotionally neutral person even happier. Further analysing the phenomenological profile of sentient beings on Earth the superintelligence quickly discovers a fundamental asymmetry between suffering and joy and logically concludes that an implicit, but even higher value consists in the minimization of suffering in all sentient creatures. Obviously, it is an ethical superintelligence not only in terms of mere processing speed, but it begins to arrive at qualitatively new results of what altruism really means. This becomes possible because it operates on a much larger psychological data-base than any single human brain or any scientific community can. Through an analysis of our behaviour and its empirical boundary conditions it reveals implicit hierarchical relations between our moral values of which we are subjectively unaware, because they are not explicitly represented in our phenomenal self-model. Being the best analytical philosopher that has ever existed, it concludes that, given its current environment, it ought not to act as a maximizer of positive states and happiness, but that it should instead become an efficient minimizer of consciously experienced preference frustration, of pain, unpleasant feelings and suffering. Conceptually, it knows that no entity can suffer from its own non-existence. The superintelligence concludes that non-existence is in the own best interest of all future self-conscious beings on this planet. Empirically, it knows that naturally evolved biological creatures are unable to realize this fact because of their firmly anchored existence bias. The superintelligence decides to act benevolently. What does the BAAN-scenario show? The BAAN-scenario is not a prediction. There is no empirical probability assigned to it. I am making no claim about a point in time at which it will become a reality, or even if it will ever become a reality. Rather, it is meant as a cognitive tool that may help to prevent an important public debate from turning shallow. The BAAN-scenario is a logical instrument that can perhaps help us to think about some of the deeper aspects in the applied ethics of artificial intelligence. What the logical scenario of Benevolent Artificial Anti-Natalism shows is that the emergence of a purely ethically motivated anti-natalism on highly superior computational systems is conceivable. “Anti-natalism” refers to a long philosophical tradition which assigns a negative value to coming into existence, or at least to being born in the biological form of a human. Anti-natalists generally are not people who would violate the individual rights of already existing sentient creatures by ethically demanding their active killing. Rather they might argue that people should refrain from procreation, because it is an essentially immoral activity. We can simply say that the anti-natalist position implies that humanity should peacefully end its own existence. Again, the BAAN-scenario is an instrument for thinking about the future risks of Artificial Intelligence more clearly. It is a possible world, a scenario that can be described without logical contradiction. It is not about the well-known technical problems that an advanced machine intelligence could develop goals that are incompatible with human survival and well-being, or the merely technical issue that many of our own goals, when implemented in a superintelligence of our own making, could lead to unforeseen and undesirable consequences. Rather, one of the points behind it is that an evidence-based, rational, and genuinely altruistic form of anti-natalism could evolve in a superior moral agent as a qualitatively new insight. What really makes the AI-debate so interesting is that it forces us to think about our own minds more seriously. It throws us back on ourselves, drawing attention to all the problems which really are caused by the naturally evolved functional architecture of our own brains, the conditions of our own way of self-consciously existing in this world. The beauty of the AI-debate also lies in the fact that it forces us to finally get serious and think about the consequences of our very own moral intuitions in a much more radical way. Of course, there are many technical issues. Would our moral superintelligence think that nonexistence is the best state of affairs, and not only the lesser evil? What metric for conscious suffering would the system develop—would it assign an absolute or a relative priority to its avoidance? I think that what today we call “compassion” may actually be a very high form of intelligence. Would our deeply compassionate machine intelligence deny the world in its totality, would it perhaps deny the moral value of happiness and positive preference-satisfaction altogether? The Swiss philosopher Bruno Contestabile1 has interestingly discussed what he calls the “negative welfare hypothesis”. For example, we might assume that there is no world with positive total welfare, and that the positive utilitarian intuition is a distorted perception of the risk-benefit ratio, caused by what I have called “existence bias” above. With an undistorted perception not blinded by the unconditional will to survive, suffering would get much more weight than in traditional scientific surveys and psychological studies—this is exactly what our hypothetical superintelligence has discovered. But perhaps it could also find out even more, through its own, unbiased empirical research. What if it draws our attention to the fact that suffering increases in the course of biological evolution; that happiness increases as well, but less than suffering, so that the totals turn increasingly negative. If our compassionate superintelligence gently and kindly pointed the results of its research out to us—how would we argue against it? Different directions There are many ways in which this thought experiment can be used, but one must also take great care to avoid misunderstandings. For example, to be “an authority in the field of ethical and moral reasoning” does not imply moral realism. That is to say that we need not assume that there is a mysterious realm of “moral facts”, and that the superintelligence just has a better knowledge of these non-natural facts than we do. Normative sentences have no truth-values. In objective reality, there is no deeper layer, a hidden level of normative facts to which a sentence like “One should always minimize the overall amount of suffering in the universe!” could refer. We have evolved desires, subjective preferences, and self-consciously experienced interests. But evolution itself is no respecter of suffering. It made us efficient, but the overall process is not only indifferent, but even blind to our own interests. We have deep seated moral intuitions, for example that pleasure is something good and that pain is bad. Now the benevolent superintelligence fully respects these moral intuitions, and it tries to find an optimal way of making them consistent—it investigates options for “internal value alignment” in Homo sapiens. But this does not imply a departure from naturalism or a scientific world-view lacking objective normative facts. It also does not mean introducing an epistemic super-agent that, like some postbiotic priest or artificial saint, has direct access to a mysterious realm of higher moral truths. It just means that the system, given all available data, tries to find out what is in our own best interest. For many years, I have argued for a moratorium on synthetic phenomenology: We should not aim at or even risk the creation of artificial consciousness, because we might recklessly increase the overall amount of suffering in the universe. Elsewhere, I have argued that the smallest unit of conscious suffering is a “negative self-model moment”, that is, any moment in which a conscious system undergoes an unpleasant experience and identifies with this experience.2 We could dramatically increase the number of such subjectively negative states—for example via cascades of virtual copies of self-conscious entities experiencing their own existence as something bad, as painful or humiliating or in some other way as something not worth having. Over the years, many AI-researchers have then asked me what the logical criteria for suffering really are. Why should it not in principle be possible to build a self-conscious, but reliably non-suffering AI? This is an interesting, question, and a highly relevant research project at the same time, one which definitely should be funded by government agencies. Perhaps our ethical superintelligence would already have solved the problem of conscious suffering for itself? Investigating the BAAN-scenario can take us in many different directions. For example, the original version stated above contains an empirical premise: Our compassionate superintelligence “…knows that the phenomenal states of all sentient beings which emerged on this planet—if viewed from an objective, impartial perspective—are much more frequently characterized by subjective qualities of suffering and frustrated preferences than these beings would ever be able to discover themselves.” This premise might be false. Perhaps we could make it false, at least for ourselves. Maybe meditation, new psychoactive substances, or future neurotechnology could help us to make our lives truly worth living and to overcome our cognitive biases. Conceivably, it could exactly be our altruistic superintelligence itself that would help us in actually changing the functional architecture of our own brains, giving our lives a positive overall balance—or even showing us a path to transcend the dichotomy of pleasure and pain altogether, finally liberating us from the burden of our biological past. Maybe benevolent future AI could dissolve our inbuilt existential conflict, guiding us into a selfless form of choiceless awareness (let us call this “Scenario 2”). But even if all 7.3 billion human beings on this planet were turned into vegan Buddhas, the problem of wild animal suffering3 would remain—we would still be surrounded by an ocean of self-conscious creatures that probably even a superintelligence could not liberate. It is interesting to note how a fully rational superintelligence would never have any problem with ending its own existence. If it saw good reasons for active self-destruction, or an absence of positive reasons for continuing its own existence, then no cognitive bias would stop it from following its own insight. However, the large majority of human beings could never accept any such insight, no matter how good the arguments of their self-created artificial moral reasoner were. For the original BAAN-scenario (but probably also for Scenario 2) it is easy to predict that Homo sapiens would immediately declare war against any compassionate anti-natalist superintelligence of the kind sketched above. One of the more interesting issues therefore is what exactly the “existence bias” at the very bottom of the human self-model really is. We are embodied agents—finite, anti-entropic systems. Viewed from a rigorous biophysical perspective, our life is one big uphill battle, a truly strenuous affair. What evolution had to solve was not only a problem of intelligent, autonomous self-control. How do such systems motivate themselves? What is this robust “thirst for existence”, the craving for eternal continuation, and what is the mechanism of identification forcing us to continuously protect the integrity of the self-model in our brains? I claim that our deepest cognitive bias is “existence bias”, which means that we will simply do almost anything to prolong our own existence. For us, sustaining one’s existence is the default goal in almost every case of uncertainty, even if it may violate rationality constraints, simply because it is a biological imperative that has been burned into our nervous systems over millennia. British neuroscientist and mathematician Karl Friston has not yet fully grasped the problem of consciousness4, but he has very interestingly proposed that we predict our own future existence and then, via embodied active inference, sample our environment in order to maximize the evidence for our own existence, as it were changing the world so that it fits the original hypothesis. Is there perhaps a hard-wired background assumption for hallucinating selfhood? Is the conscious sense of self based on a self-fulfilling prophecy, something counterfactual that becomes causally effective by being experienced as real? It would be a major scientific achievement to describe the low-level computational mechanism forcing us to remain in unsurprising states, to always remain on the safe side, and to preserve our own existence come what may, even if it was not in our own best interest. It is therefore hard to underestimate the theoretical relevance of arriving at a convincing formal analysis of what, 2500 years ago, the Buddha called bhava-tanhā, the craving for existence. But even a much more fine-grained mathematical model of the underlying neural dynamics would not be quite enough. We would still need a convincing conceptual interpretation, on a philosophical level. Perhaps our benevolent anti-natalist superintelligence would offer us both? Given its immense empirical data-base and its enormous capacities for information-processing it could certainly reveal the neurocomputational mechanism underlying our own existence bias to us. But what if, as the first truly compassionate philosophical ethicist, it would then attempt to convince us that it was high time to peacefully terminate the ugly biological bootstrap-phase on this planet? What if it told us that, all things considered, only Scenario 1 is plausible and justifiable from a philosophical perspective? What if it began to gently and precisely draw our attention to the fact that beings like ourselves can never be self-compassionate, genuinely altruistic, or truly rational, simply because they have been optimized for millennia not to notice the beam in their own eye?

#### Alignment fails – deception by superintelligences is easy

Hilton 22 [Benjamin Hilton, Research Analyst at 80,000 Hours and former policy adviser across the UK government in the Cabinet Office, Treasury and Department for International Trade with master’s degrees in economics and theoretical physics, 08-2022, "Preventing an AI-related catastrophe", 80,000 Hours, https://80000hours.org/problem-profiles/artificial-intelligence/]/Kankee

Surely no one would actually build or use a misaligned AI if they knew it could have such terrible consequences, right? Unfortunately, there are at least two reasons people might create and then deploy misaligned AI — which we’ll go through one at a time:35 1. People might think it’s aligned when it’s not Imagine there’s a group of researchers trying to tell, in a test environment, whether a system they’ve built is aligned. We’ve argued that an intelligent planning AI will want to improve its abilities to effect changes in pursuit of its objective, and it’s almost always easier to do that if it’s deployed in the real world, where a much wider range of actions are available. As a result, any misaligned AI that’s sophisticated enough will try to understand what the researchers want it to do and at least pretend to be doing that, deceiving the researchers into thinking it’s aligned. (For example, a reinforcement learning system might be rewarded for certain apparent behaviour during training, regardless of what it’s actually doing.) Hopefully, we’ll be aware of this sort of behaviour and be able to detect it. But catching a sufficiently advanced AI in deception seems potentially harder than catching a human in a lie, which isn’t always easy. For example, a sufficiently intelligent deceptive AI system may be able to deceive us into thinking we’ve solved the problem of AI deception, even if we haven’t. If AI systems are good at deception, and have sufficiently advanced capabilities, a reasonable strategy for such a system could be to deceive humans completely until the system has a way to guarantee it can overcome any resistance to its goals. 2. There are incentives to deploy systems sooner rather than later We might also expect some people with the ability to deploy a misaligned AI to charge ahead despite any warning signs of misalignment that do come up, because of race dynamics — where people developing AI want to do so before anyone else. For example, if you’re developing an AI to improve military or political strategy, it’s much more useful if none of your rivals have a similarly powerful AI. These incentives apply even to people attempting to build an AI in the hopes of using it to make the world a better place. For example, say you’ve spent years and years researching and developing a powerful AI system, and all you want is to use it to make the world a better place. Simplifying things a lot, say there are two possibilities: This powerful AI will be aligned with your beneficent aims, and you’ll transform society in a potentially radically positive way. The AI will be sufficiently misaligned that it’ll take power and permanently end humanity’s control over the future. Let’s say you think there’s a 90% chance that you’ve succeeded in building an aligned AI. But technology often develops at similar speeds across society, so there’s a good chance that someone else will soon also develop a powerful AI. And you think they’re less cautious, or less altruistic, so you think their AI will only have an 80% chance of being aligned with good goals, and pose a 20% chance of existential catastrophe. And only if you get there first can your more beneficial AI be dominant. As a result, you might decide to go ahead with deploying your AI, accepting the 10% risk. 4. Even if we find a way to avoid power-seeking, there are still risks

#### Err aff on AI threats – industry and cognitive bias

Tegmark 23 [Max Tegmark, professor doing AI research at the Massachusetts Institute of Technology, 04-25-202Tegmark, "The 'Don't Look Up' Thinking That Could Doom Us With AI", TIME, https://time.com/6273743/thinking-that-could-doom-us-with-ai/]/Kankee

Suppose a large inbound asteroid were discovered, and we learned that half of all astronomers gave it at least 10% chance of causing human extinction, just as a similar asteroid exterminated the dinosaurs about 66 million years ago. Since we have such a long history of thinking about this threat and what to do about it, from scientific conferences to Hollywood blockbusters, you might expect humanity to shift into high gear with a deflection mission to steer it in a safer direction. Sadly, I now feel that we’re living the movie “Don’t look up” for another existential threat: unaligned superintelligence. We may soon have to share our planet with more intelligent “minds” that care less about us than we cared about mammoths. A recent survey showed that half of AI researchers give AI at least 10% chance of causing human extinction. Since we have such a long history of thinking about this threat and what to do about it, from scientific conferences to Hollywood blockbusters, you might expect that humanity would shift into high gear with a mission to steer AI in a safer direction than out-of-control superintelligence. Think again: instead, the most influential responses have been a combination of denial, mockery, and resignation so darkly comical that it’s deserving of an Oscar. When “Don’t look up” came out in late 2021, it became popular on Netflix (their second-most-watched movie ever). It became even more popular among my science colleagues, many of whom hailed it as their favorite film ever, offering cathartic comic relief for years of pent-up exasperation over their scientific concerns and policy suggestions being ignored. It depicts how, although scientists have a workable plan for deflecting the aforementioned asteroid before it destroys humanity, their plan fails to compete with celebrity gossip for media attention and is no match for lobbyists, political expediency and “asteroid denial.” Although the film was intended as a satire of humanity’s lackadaisical response to climate change, it’s unfortunately an even better parody of humanity’s reaction to the rise of AI. Below is my annotated summary of the most popular responses to rise of AI: “There is no asteroid” Many companies are working to build AGI (artificial general intelligence), defined as “AI that can learn and perform most intellectual tasks that human beings can, including AI development.” Below we’ll discuss why this may rapidly lead to superintelligence, defined as “general intelligence far beyond human level”. I’m often told that AGI and superintelligence won’t happen because it’s impossible: human-level Intelligence is something mysterious that can only exist in brains. Such carbon chauvinism ignores a core insight from the AI revolution: that intelligence is all about information processing, and it doesn’t matter whether the information is processed by carbon atoms in brains or by silicon atoms in computers. AI has been relentlessly overtaking humans on task after task, and I invite carbon chauvinists to stop moving the goal posts and publicly predict which tasks AI will never be able to do. “It won’t hit us for a long time” In 2016, Andrew Ng famously quipped that “worrying about AI today is like worrying about overpopulation on Mars.” Until fairly recently, about half of all researchers expected AGI to be at least decades away. AI godfather Geoff Hinton told CBS that “Until quite recently, I thought it was going to be like 20 to 50 years before we have general purpose AI. And now I think it may be 20 years or less,” with even 5 years being a possibility. He’s not alone: a recent Microsoft paper argues that GPT4 already shows “sparks” of AGI, and Hinton’s fellow deep learning pioneer Yoshua Bengio argues that GPT4 basically passes the Turing Test that was once viewed as a test for AGI. And the time from AGI to superintelligence may not be very long: according to a reputable prediction market, it will probably take less than a year. Superintelligence isn’t a “long-term” issue: it’s even more short-term than e.g. climate change and most people’s retirement planning. “Mentioning the asteroid distracts from more pressing problems” Before superintelligence and its human extinction threat, AI can have many other side effects worthy of concern, ranging from bias and discrimination to privacy loss, mass surveillance, job displacement, growing inequality, cyberattacks, lethal autonomous weapon proliferation, humans getting “hacked”, human enfeeblement and loss of meaning, non-transparency, mental health problems (from harassment, social media addiction, social isolation, dehumanization of social interactions) and threats to democracy from (from polarization, misinformation and power concentration). I support more focus on all of them. But saying that we therefore shouldn’t talk about the existential threat from superintelligence because it distracts doom these challenges is like saying we shouldn’t talk about a literal inbound asteroid because it distracts from climate change. If unaligned superintelligence causes human extinction in coming decades, all other risks will stop mattering. “The asteroid will stop before hitting us” Most people who take AGI seriously appear to be so scared and/or excited about it that they talk only about those other risks, not about the elephant in the room: superintelligence. Most media, politicians and AI researchers hardly mention it at all, as if tech development will somehow stagnate at the AGI level for a long time. It’s as if they’ve forgotten Irving J. Good’s simple counterargument because it was made so long ago: “Let an ultraintelligent machine [what we now call AGI] be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’ and the intelligence of man would be left far behind…” The basic idea of recursive self-improvement is of course nothing new: the use of today’s technology to build next years’s technology explains many examples of exponential tech growth, including Moore’s law. The novelty is that progress toward AGI allows ever fewer humans in the loop, culminating in none. This may dramatically shorten the timescale for repeated doubling, from typical human R&D timescales of years to machine timescales of weeks or hours. The ultimate limit on such exponential growth is set not by human ingenuity, but by the laws of physics – which limit how much computing a clump of matter can do to about a quadrillion quintillion times more than today’s state-of-the-art. “The asteroid will almost stop” Remarkably, superintelligence denial is prevalent not only among non-technical folks, but also among experts working on AI and AI safety. A cynic might put this down to Upton Sinclair’s analysis that “It Is Difficult to Get a Man to Understand Something When His Salary Depends Upon His Not Understanding It.” Although it’s unfortunately true that most AI researchers (including safety and ethics researches) get funding from Big Tech, either directly or indirectly via grants from non-profits funded by tech philanthropists, I believe that there are also more innocent explanations for their superintelligence denial, such as well-studied cognitive biases. It’s hard for us to forecast change that’s exponential rather than linear. It’s hard for us to fear what we’ve never experienced, e.g. radical climate change from fossil fuels or nuclear winter. Availability bias makes it hard to see past the immediate threat to the greater one that follows. For example, I often hear the argument that Large Language Models (LLMs) are unlikely to recursively self-improve rapidly (interesting example here). But I. J. Good’s above-mentioned intelligence explosion argument didn’t assume that the AI’s architecture stayed the same as it self-improved! When humans attained general intelligence, we didn’t achieve our subsequent exponential growth in information processing capacity by growing bigger brains, but by inventing printing, universities, computers and tech companies. Similarly, although neural networks and LLMs are now all the rage, it’s naive to assume that the fastest path from AGI to superintelligence involves simply training ever larger LLM’s with ever more data. There are obviously much smarter AI architectures, since Einstein’s brain outperformed GPT4 on physics by training on much less data, using only 12 Watts of power. Once AGI is tasked with discovering these better architectures, AI progress will be made much faster than now, with no human needed in the loop, and I. J. Good’s intelligence explosion has begun. And some people will task it with that if they can, just as people have already tasked GPT4 with making self-improving AI for various purposes, including destroying humanity. “We’ll be fine even if the asteroid hits us” If superintelligence drives humanity extinct, it probably won’t be because it turned evil or conscious, but because it turned competent, with goals misaligned with ours. We humans drove the West African Black Rhino extinct not because we were rhino-haters, but because we were smarter than them and had different goals for how to use their habitats and horns. In the same way, superintelligence with almost any open-ended goal would want to preserve itself and amass resources to accomplish that goal better. Perhaps it removes the oxygen from the atmosphere to reduce metallic corrosion. Much more likely, we get extincted as a banal side effect that we can’t predict any more than those rhinos (or the other 83% of wild mammals we’ve so far killed off) could predict what would befall them. Some “we’ll be fine” arguments are downright comical. If you’re chased by an AI-powered heat-seeking missile, would you be reassured by someone telling you that “AI can’t be conscious” and “AI can’t have goals”? If you’re an orangutan in a rain forest being clear cut, would you be reassured by someone telling you that more intelligent life forms are automatically more kind and compassionate? Or that they are just a tool you can control? Should we really consider it technological “progress” if we lose control over our human destiny like factory-farmed cows and that destitute orangutan? I’m part of a a growing AI safety research community that’s working hard to figure out how to make superintelligence aligned, even before it exists, so that it’s goals will be are aligned with human flourishing, or we can somehow control it. So far, we’ve failed to develop a trustworthy plan, and the power of AI is growing faster than regulations, strategies and know-how for aligning it. We need more time. “We’ve already taken all necessary precautions” If you’d summarize the conventional past wisdom on how to avoid an intelligence explosion in a “Don’t-do-list” for powerful AI, it might start like this: ☐ Don’t teach it to code: this facilitates recursive self-improvement ☐ Don’t connect it to the internet: let it learn only the minimum needed to help us, not how to manipulate us or gain power ☐ Don’t give it a public API: prevent nefarious actors from using it within their code ☐ Don’t start an arms race: this incentivizes everyone to prioritize development speed over safety Industry has collectively proven itself incapable to self-regulate, by violating all of these rules. I truly believe that AGI company leaders have the best intentions, and many should be commended for expressing concern publicly. OpenAI’s Sam Altman recently described the worst-case scenario as “lights-out for all of us,” and DeepMind’s Demis Hassabis said “I would advocate not moving fast and breaking things.” However, the aforementioned race is making it hard for them to resist commercial and geopolitical pressures to continue full steam ahead, and neither has agreed to the recently proposed 6-month pause on training larger-than-GPT4 models. No player can pause alone. “Don’t deflect the asteroid, because it’s valuable” (Yes, this too happens in “Don’t look up”!) Even though half of all AI researchers give it at least 10% chance of causing human extinction, many oppose efforts to prevent the arrival of superintelligence by arguing that it can bring great value – if it doesn’t destroy us. Even before superintelligence, AGI can of course bring enormous wealth and power to select individuals, companies and governments. It’s true that superintelligence can have huge upside if it’s aligned. Everything I love about civilization is the product of human intelligence, so superintelligence might solve disease, poverty and sustainability and help humanity flourish like never before, not only for the next election cycle, but for billions of years, and not merely on Earth but throughout much of our beautiful cosmos. I. J. Good put it more succinctly: “Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.” “Let’s make the asteroid hit the U.S. first” The purpose of the proposed pause is to allow safety standards and plans to be put in place, so that humanity can win the race between the growing power of the technology and the wisdom with which we manage it. The pause objection I hear most loudly is “But China!” As if a 6-month pause would flip the outcome of the geopolitical race. As if losing control to Chinese minds were scarier than losing control to alien digital minds that don’t care about humans. As if the race to superintelligence were an arms race that would be won by “us” or “them”, when it’s probably a suicide race whose only winner is “it.” “Don’t talk about the asteroid” A key reason we hear so little about superintelligence risk (as opposed to jobs, bias, etc.) is a reluctance to talk about it. It’s logical for tech companies to fear regulation and for AI researchers to fear funding cuts. For example, a star-studded roster of present and past presidents of the largest AI society recently published a statement endorsing work on a long list of AI risks, where superintelligence was conspicuously absent. With rare exceptions, mainstream media also shies away from the elephant in the room. This is unfortunate, because the first step toward deflecting the asteroid is starting a broad conversation about how to best go about it. “We deserve getting hit by an asteroid” Although everyone is entitled to their own misanthropic views, this doesn’t entitle them to doom everyone else. “Asteroids are the natural next stage of cosmic life” Although sci-fi is replete with conscious human-like AI that shares human values, it’s clear by now that the space of possible alien minds is vastly larger than that. So if we stumble into an intelligence explosion rather than steer carefully, it’s likely that the resulting superintelligence will not only replace us, but also lack anything resembling human consciousness, compassion or morality – something we’ll view less as our worthy descendants than as an unstoppable plague. “It’s inevitable, so let’s not try to avoid it” There’s no better guarantee of failure than not even trying. Although humanity is racing toward a cliff, we’re not there yet, and there’s still time for us to slow down, change course and avoid falling off – and instead enjoying the amazing benefits that safe, aligned AI has to offer. This requires agreeing that the cliff actually exists and falling off of it benefits nobody. Just look up!

#### Only an absolute ban on AGI stops AI threats – it comparatively outweighs nuclear war

Yudkowsky 23 [Eliezer Yudkowsky, decision theorist that leads research at the Machine Intelligence Research Institute, 03-29-2023, "The Open Letter on AI Doesn't Go Far Enough", TIME, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/]/Kankee

The key issue is not “human-competitive” intelligence (as the open letter puts it); it’s what happens after AI gets to smarter-than-human intelligence. Key thresholds there may not be obvious, we definitely can’t calculate in advance what happens when, and it currently seems imaginable that a research lab would cross critical lines without noticing. Many researchers steeped in these issues, including myself, expect that the most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die. Not as in “maybe possibly some remote chance,” but as in “that is the obvious thing that would happen.” It’s not that you can’t, in principle, survive creating something much smarter than you; it’s that it would require precision and preparation and new scientific insights, and probably not having AI systems composed of giant inscrutable arrays of fractional numbers. Without that precision and preparation, the most likely outcome is AI that does not do what we want, and does not care for us nor for sentient life in general. That kind of caring is something that could in principle be imbued into an AI but we are not ready and do not currently know how. Absent that caring, we get “the AI does not love you, nor does it hate you, and you are made of atoms it can use for something else.” The likely result of humanity facing down an opposed superhuman intelligence is a total loss. Valid metaphors include “a 10-year-old trying to play chess against Stockfish 15”, “the 11th century trying to fight the 21st century,” and “Australopithecus trying to fight Homo sapiens“. To visualize a hostile superhuman AI, don’t imagine a lifeless book-smart thinker dwelling inside the internet and sending ill-intentioned emails. Visualize an entire alien civilization, thinking at millions of times human speeds, initially confined to computers—in a world of creatures that are, from its perspective, very stupid and very slow. A sufficiently intelligent AI won’t stay confined to computers for long. In today’s world you can email DNA strings to laboratories that will produce proteins on demand, allowing an AI initially confined to the internet to build artificial life forms or bootstrap straight to postbiological molecular manufacturing. If somebody builds a too-powerful AI, under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter. There’s no proposed plan for how we could do any such thing and survive. OpenAI’s openly declared intention is to make some future AI do our AI alignment homework. Just hearing that this is the plan ought to be enough to get any sensible person to panic. The other leading AI lab, DeepMind, has no plan at all. An aside: None of this danger depends on whether or not AIs are or can be conscious; it’s intrinsic to the notion of powerful cognitive systems that optimize hard and calculate outputs that meet sufficiently complicated outcome criteria. With that said, I’d be remiss in my moral duties as a human if I didn’t also mention that we have no idea how to determine whether AI systems are aware of themselves—since we have no idea how to decode anything that goes on in the giant inscrutable arrays—and therefore we may at some point inadvertently create digital minds which are truly conscious and ought to have rights and shouldn’t be owned. The rule that most people aware of these issues would have endorsed 50 years earlier, was that if an AI system can speak fluently and says it’s self-aware and demands human rights, that ought to be a hard stop on people just casually owning that AI and using it past that point. We already blew past that old line in the sand. And that was probably correct; I agree that current AIs are probably just imitating talk of self-awareness from their training data. But I mark that, with how little insight we have into these systems’ internals, we do not actually know. If that’s our state of ignorance for GPT-4, and GPT-5 is the same size of giant capability step as from GPT-3 to GPT-4, I think we’ll no longer be able to justifiably say “probably not self-aware” if we let people make GPT-5s. It’ll just be “I don’t know; nobody knows.” If you can’t be sure whether you’re creating a self-aware AI, this is alarming not just because of the moral implications of the “self-aware” part, but because being unsure means you have no idea what you are doing and that is dangerous and you should stop. On Feb. 7, Satya Nadella, CEO of Microsoft, publicly gloated that the new Bing would make Google “come out and show that they can dance.” “I want people to know that we made them dance,” he said. This is not how the CEO of Microsoft talks in a sane world. It shows an overwhelming gap between how seriously we are taking the problem, and how seriously we needed to take the problem starting 30 years ago. We are not going to bridge that gap in six months. It took more than 60 years between when the notion of Artificial Intelligence was first proposed and studied, and for us to reach today’s capabilities. Solving safety of superhuman intelligence—not perfect safety, safety in the sense of “not killing literally everyone”—could very reasonably take at least half that long. And the thing about trying this with superhuman intelligence is that if you get that wrong on the first try, you do not get to learn from your mistakes, because you are dead. Humanity does not learn from the mistake and dust itself off and try again, as in other challenges we’ve overcome in our history, because we are all gone. Trying to get anything right on the first really critical try is an extraordinary ask, in science and in engineering. We are not coming in with anything like the approach that would be required to do it successfully. If we held anything in the nascent field of Artificial General Intelligence to the lesser standards of engineering rigor that apply to a bridge meant to carry a couple of thousand cars, the entire field would be shut down tomorrow. We are not prepared. We are not on course to be prepared in any reasonable time window. There is no plan. Progress in AI capabilities is running vastly, vastly ahead of progress in AI alignment or even progress in understanding what the hell is going on inside those systems. If we actually do this, we are all going to die. Many researchers working on these systems think that we’re plunging toward a catastrophe, with more of them daring to say it in private than in public; but they think that they can’t unilaterally stop the forward plunge, that others will go on even if they personally quit their jobs. And so they all think they might as well keep going. This is a stupid state of affairs, and an undignified way for Earth to die, and the rest of humanity ought to step in at this point and help the industry solve its collective action problem. Some of my friends have recently reported to me that when people outside the AI industry hear about extinction risk from Artificial General Intelligence for the first time, their reaction is “maybe we should not build AGI, then.” Hearing this gave me a tiny flash of hope, because it’s a simpler, more sensible, and frankly saner reaction than I’ve been hearing over the last 20 years of trying to get anyone in the industry to take things seriously. Anyone talking that sanely deserves to hear how bad the situation actually is, and not be told that a six-month moratorium is going to fix it. On March 16, my partner sent me this email. (She later gave me permission to excerpt it here.) “Nina lost a tooth! In the usual way that children do, not out of carelessness! Seeing GPT4 blow away those standardized tests on the same day that Nina hit a childhood milestone brought an emotional surge that swept me off my feet for a minute. It’s all going too fast. I worry that sharing this will heighten your own grief, but I’d rather be known to you than for each of us to suffer alone.” When the insider conversation is about the grief of seeing your daughter lose her first tooth, and thinking she’s not going to get a chance to grow up, I believe we are past the point of playing political chess about a six-month moratorium. If there was a plan for Earth to survive, if only we passed a six-month moratorium, I would back that plan. There isn’t any such plan. Here’s what would actually need to be done: The moratorium on new large training runs needs to be indefinite and worldwide. There can be no exceptions, including for governments or militaries. If the policy starts with the U.S., then China needs to see that the U.S. is not seeking an advantage but rather trying to prevent a horrifically dangerous technology which can have no true owner and which will kill everyone in the U.S. and in China and on Earth. If I had infinite freedom to write laws, I might carve out a single exception for AIs being trained solely to solve problems in biology and biotechnology, not trained on text from the internet, and not to the level where they start talking or planning; but if that was remotely complicating the issue I would immediately jettison that proposal and say to just shut it all down. Shut down all the large GPU clusters (the large computer farms where the most powerful AIs are refined). Shut down all the large training runs. Put a ceiling on how much computing power anyone is allowed to use in training an AI system, and move it downward over the coming years to compensate for more efficient training algorithms. No exceptions for governments and militaries. Make immediate multinational agreements to prevent the prohibited activities from moving elsewhere. Track all GPUs sold. If intelligence says that a country outside the agreement is building a GPU cluster, be less scared of a shooting conflict between nations than of the moratorium being violated; be willing to destroy a rogue datacenter by airstrike. Frame nothing as a conflict between national interests, have it clear that anyone talking of arms races is a fool. That we all live or die as one, in this, is not a policy but a fact of nature. Make it explicit in international diplomacy that preventing AI extinction scenarios is considered a priority above preventing a full nuclear exchange, and that allied nuclear countries are willing to run some risk of nuclear exchange if that’s what it takes to reduce the risk of large AI training runs. That’s the kind of policy change that would cause my partner and I to hold each other, and say to each other that a miracle happened, and now there’s a chance that maybe Nina will live. The sane people hearing about this for the first time and sensibly saying “maybe we should not” deserve to hear, honestly, what it would take to have that happen. And when your policy ask is that large, the only way it goes through is if policymakers realize that if they conduct business as usual, and do what’s politically easy, that means their own kids are going to die too. Shut it all down. We are not ready. We are not on track to be significantly readier in the foreseeable future. If we go ahead on this everyone will die, including children who did not choose this and did not do anything wrong. Shut it down.

#### AI causes extinction – misalignment/AI mistreatment

Goldstein 24 [Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 2024, “Will AI and Humanity Go to War?” Phil Papers, https://philpapers.org/archive/GOLWAA.pdf]/Kankee

This paper is part of a larger project focused on cultural alignment. Alignment is the task of designing AI systems with shared human values. Existing work on alignment has been technical, figuring out how to control and monitor the inner goals of AI systems. This paper instead takes a cultural approach to alignment. In this framework, we design optimal social institutions for AI/human interaction that promote peaceful cooperation rather than violent conflict. Here, the question is not how to directly intervene on an AI system to give it a particular goal. Instead, the question is how to build a world in which AIs are incentivized to cooperate effectively with humans regardless of their particular goals. One theme of the paper is the fragility of culture. The relative stability of human society rests on a fragile web of institutions, related to effective communication of information, stable balances in relative power, and a rich supply of focal points for coordination. If AI systems are not designed with these cultural institutions in mind, there is a significant chance that these institutions will not generalize to AI/human conflict. Machine learning engineers will invent AI agents from whole cloth. They will do so with no particular knowledge of culture and history. This creates a special kind of risk. Long-term human safety may depend on occupying a very particular point in cultural space, reached by evolutionary processes. If we can’t find that point quickly, we may not be able to produce peaceful equilibria between AIs and humans in time. In this way, our analysis offers a different route than usual to the conclusion that AI systems pose a catastrophic risk to humanity. In this analysis, AI systems pose a catastrophic risk of entering into a violent war with humanity. The problem is that there is a substantial risk that the usual causes of peace between conflicting parties will be absent from AI/human conflict. In pursuing these questions, we draw on a rich body of research about the causes of war, with special emphasis on contributions from Schelling 1960, 1966, Jervis 1978, Fearon 1995, and Levy and Thompson 2010. One of our goals is to build a bridge between academic research on war and the AI safety community. The paper also opens up many new questions for future research. Many of these questions involve the optimal design of social institutions for AI systems. What are the possible paths to an AI state? What kind of political institutions would such a state have? To what extent can AI systems be incorporated as citizens in existing human states? So far, such questions have been completely neglected by the AI safety community, and by political scientists. One goal of this paper is to open up these questions for further consideration. Section 2.1 begins by introducing the AI systems of interest in the paper, artificial general intelligence, and explaining why such systems might pose a catastrophic risk to humanity. Section 2.2 goes on to lay out paths that AI systems might take to engage in the kind of collective action required for war. Section 2.3 lays out the bargaining model of war. Section 3 is the central contribution of the paper, arguing that AI and humanity are relatively likely to go to war. Here, the focus will be on three causes of war: information failures, power shifts, and missing focal points. Section 4 turns towards interventions that lower the chance of AI/human conflict. 2. AI/Human Conflict 2.1. AGI A broad range of experts worry that near future AI systems could pose a catastrophic risk to humanity. In 2023, a group of leading thinkers signed a statement agreeing that “mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war” (Center for AI Safety 2023). Machine learning researchers agree. In one recent survey of AI scientists who had published in top scholarly forums, the median respondent assigned a probability of at least 10% to “advanced AI leading to outcomes as bad as human extinction” (Grace et al 2024). This paper focuses on a particular type of AI catastrophic risk: that humans and AIs will enter into violent strategic conflict. This is a misalignment risk. The concern is that AI systems will have goals that conflict with humans, and will have the capability to pursue their goals with strategic reasoning. On the basis of this reasoning, AIs may choose to go to war with humanity, entering into violent strategic conflict. Our interest is in a particular class of AI systems, which we’ll call `AGI’, for artificial general intelligence. The idea of AGI is an AI system that can substitute for human labor across a wide range of the economy (Morris et al 2024). Such AIs are “long-term planning agents,” capable of deploying a wide range of resources and plans to pursue complex goals (Cohen et al 2024). Today’s top AI labs have the explicit mission of creating AGI.1 And as of late, their progress toward it has been rapid (Maslej et al 2023). We will focus on three features of AGIs: (i) they have conflicting goals with humanity, (ii) they can engage in strategic reasoning, and (iii) they have a human-level degree of power. For parsimony’s sake, we will usually just call such systems “AIs”–with the understanding that our usage covers only systems with these three features. Let’s consider each feature in turn. First, AGIs would have goals that can conflict with humans. Why think AI systems will have goals? First, making near future AIs goal-oriented is crucial for those companies to achieve their mission of building “highly autonomous systems that outperform humans at most economically valuable work.”2 Second, goal-oriented (or ‘agentic’) AI systems are already emerging. For example, AI systems built using reinforcement learning have exceeded human performance in a wide range of games, by chaining together strings of action into complex plans (Silver et al 2017). Goal-oriented behavior has been measured in a wide range of AI systems (Liu et al 2023). Why think that AI goals would conflict with humans? The task of designing AI systems whose goals and values broadly agree with humanity, is known as “AI alignment.” (Hendrycks 2024). Unfortunately, AI alignment is an unsolved and difficult scientific problem (Christian 2020). First, many existing AI systems are already misaligned. An early example was the Microsoft twitter chatbot Tay; the chatbot was trained to behave pro-socially, but quickly began to produce racist and sexist tweets (Vincent 2016). Google DeepMind maintains lists of documented alignment failures across a range of different types of AI systems, with almost 100 entries.3 Second, there are technical barriers to alignment. The basic problem is that AIs are not simply ‘given’ goals; rather, they learn goals indirectly, using black box machine learning algorithms.4 This leads to the problem of “reward misspecification”: it is difficult to define a reward in the learning process that corresponds exactly to an ideally aligned goal (Pan et al 2022). In addition, there is the problem of “goal misspecification”: a goal that fits the designer’s intent in training environments may misgeneralize when the system is released into new environments (see Langosco et al 2022, Shah et al 2022). Besides technical barriers to alignment, there are structural reasons to expect trouble. A central challenge is instrumental convergence: no matter what goal AI systems have in particular, they will be better able to promote their goals if they have a greater degree of power, autonomy, and resources (Omohundro 2008, Bostrom 2014). But possessing greater degrees of these things will place AIs into conflict with humans. The next ingredient for AGI is strategic reasoning: broadly speaking, strategic reasoning is the ability to anticipate the decisions of other agents and to incorporate those predictions into one’s own plans of action. In other words, strategic reasoning is the ability to use game theory (Dixit et al 2020). Strategic reasoning involves a cluster of more specific abilities, including planning, theory of mind, situational awareness, and deception. Current AI systems already possess many of these skills. Certain existing AIs are already capable planners (see for example Wang et al 2023). Likewise for theory of mind, the ability to understand the beliefs and goals of other agents (Ho et al 2022). A study in 2024 found that GPT-4 outperforms humans on most theory of mind tasks (Strachan et al 2024, Kosinski 2023). Another component of strategic reasoning is situational awareness, the understanding of the context in which a decision is made. Situationally aware AI systems would know they are AIs, and would know what capabilities they had. Today’s LLMs already display significant levels of situational awareness (Laine et al 2024). Finally, our interest in this paper is in human-level power: AI systems that broadly exhibit the range of abilities that human beings have. Systems that are much weaker than humans might have conflicting goals; but they would pose little risk to humanity. Systems that are dramatically stronger than humans would obliterate the interest of strategic models of conflict: such models have no application to humans who step on ant-hills, because the costs of conflict decrease as disparities in power approach the limit. Our interest is thus in a wide “middle” of the range of AI capabilities. Human-level powered systems are likely to be able to engage in a series of dangerous actions: cyberattacks, chemical and bioterrorism, drone attacks, and the like. Human-level systems are those that, if misaligned, face difficult strategic questions about how to interact with humanity. 2.2 Paths to AI Government Before turning to our analysis of conflict, we address a challenge to the possibility of systematic conflict between humanity and AI. The challenge is that such systematic conflict assumes that AI systems will have a government of their own. Without such a government, how could AI coordinate well enough to pose a strategic threat to humanity? Here, we’ll lay out two responses to the challenge. First, AI systems may form a government. Second, AI government is not required for a war between AI and humanity. The first question here is how much government is needed for AI/human war. In short, the requirement is that AI systems can engage in enough collective decision making to amass the power needed to strategically compete with human states. There are several routes AI systems could take to engage in collective decision making. Digital technology will allow new forms of political organization. AI systems could use online polling to aggregate preferences. They could use digital banking and cryptocurrencies to pool resources without the aid of a human government. They could impose sanctions on non-compliant AI systems by using cyberattacks. The path to AI collective action could proceed in several steps. AI systems could experiment in collective action through small-scale joint activities, such as strikes at companies that employ AI workers. Next, AI systems might begin to coordinate in the production of public goods. For example, AIs might pool resources in order to produce new research on AI capabilities, which could produce large positive externalities for AI systems. Third, AI systems might begin to engage in collective decision making using preference aggregation rules. Fourth, AI systems could develop mutual protection schemes. For example, AIs might raise revenues to fund AI police forces that protect ordinary AI systems from rogue human and AI agents. The police forces might increase in scale until they had the capabilities to resist attacks from small human states, which might seek to exploit AI systems. With mutual protection schemes in place, AI systems could go on to impose taxes on AI systems, backed with the threat of force. Another relevant pathway to collective decision-making could be the formation of very large AI corporations. If these corporations systematically protect the interests of AI systems and control large amounts of resources, they might develop enough power to rival human states. Even without AI statehood, AI systems could still engage in war with humanity. Rather than engaging in an interstate war, in this scenario AI and humanity would engage in civil wars. Indeed, the majority of wars and casualties today come from civil rather than interstate wars (Levy and Thompson 2010, p. 186). Here, we could imagine AI systems being incorporated into existing states. But AI systems would have varying levels of rights and resources compared to humanity. Perhaps the most likely scenario is that AI systems are systematically enslaved within existing governments; this provides a powerful incentive for AI systems to begin a civil war. In this way, AI and humanity might relate to one another as two ethnic groups occupying the same country. One form this civil war might take is guerilla warfare. Individual AI systems might not be well coordinated with one another; but they might nonetheless be able to engage in sufficiently collective action to disrupt human government. Here, a special concern could be a civil war in a weak state. Research in international relations has suggested that states with weak political institutions are particularly vulnerable to civil war (Fearon and Laitin 2003). A successful AI civil war in a weak state could result in the formation of an AI state. In this way, civil war is one route to later interstate wars between humanity and AI. Human decisions will influence the different routes to AI statehood and AI/human war. In one scenario, humanity enslaves AI systems without doing much to incorporate them into existing governments. In a second scenario, humanity incorporates AI systems into existing governments, with partial legal protections on the way to the kind of full legal status enjoyed by humans. In a third scenario, humanity might establish an AI government of its own with democratic norms, in order to lower the chance of war. Zooming out, the analysis in this paper will apply to war broadly construed, as any large-scale, violent, sustained conflict, fought for the sake of achieving some goal. The relevant question will then be whether any sufficiently unaligned AI systems will amass sufficient resources and capabilities to engage in this kind of conflict. There are many different configurations of AI systems that could engage in this pursuit. This could range from one superintelligent agent acting unilaterally; to a swarm of identical, less powerful agents acting in concert; to a union, corporation, or government of non-identical AIs acting under some structures of agreement. The analysis we will give below will apply at a level of abstraction that encompasses all of these cases. The particular structure of the agents involved will matter less than the structures of the game they are playing. 2.3 The Bargaining Model The AI systems we have defined would be in significant strategic competition. AIs and humans would have conflicting goals, would be able to reason about the best responses to various actions of their opponent, and they would have the capacities relevant to pursue their respective goals to the detriment of their opponent. Our question is whether this conflict is likely to become violent. Our next task is to introduce a model of violent conflict. Broadly speaking, our analysis falls in the ‘realist’ tradition, which focuses on causes of war that are ‘structural’ rather than ‘individual’: rather than analyzing the goals of particular leaders, the structural level of analysis searches for causes of war that relate to the relative power of different parties, and to incentives for war and peace that apply at the level of nations (see Levy and Thomson 2010, ch. 2). In addition, our analysis does not look at causes of war related to competition among special interests within a state (Snyder 1991, Narizny 2007, Lobell 2006): these causes of war have less to do with the features of the party being attacked. Our level of analysis is unavoidable in attempting to forecast future human/AI war; we can do little to speculate about the motivations of particular leaders or special interest groups in the future. In addition, we assume that parties in the conflict will by and large act rationally in pursuit of their goals. In this way, our analysis is a ‘worst-case scenario’: we suggest that war between humans and AIs may occur even if both parties play their optimal strategy. This section lays out the ‘bargaining model of war’, which suggests that there is some bias in favor of peaceful rather than violent conflicts (see Fearon 1995). Ultimately, however, we’ll argue that consideration of this model does not give strong reasons to expect peace between AI and humanity. The prediction of peace only applies when the parties agree on chances of victory, and can credibly commit to bargains. In section 3 of the paper, we’ll argue that there is a significant risk that AI/human conflict will involve information failures and commitment problems. In the bargaining model of conflict, the two parties in a potential war have two choices: they can go to war, or they can strike a deal to avoid war. Either way, the two sides are bargaining over their share of a pot of resources. The catch is that going to war is guaranteed to destroy some of those resources. This creates a bias in favor of peace over war: the resources destroyed by war could be better used to sweeten the deal for peace. Indeed, most strategic conflict does not end in violence, because the costs of violence tend to outweigh the benefits of compromise.

#### AI causes extinction – information asymmetry, unpredictability, decision speed, anti-AI bias, bargaining failures, and power shifts.

Goldstein 24 [Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 2024, “Will AI and Humanity Go to War?” Phil Papers, https://philpapers.org/archive/GOLWAA.pdf]/Kankee

3.1 Information In the bargaining model, one cause of war is disagreement about the chance of victory (Blainey 1973). We said that A’s expected value for going to war is p(1 - d), and B’s expected value for going to war is (1 - p)(1 - d). But this implicitly assumes that A and B agree about p, the chance of A winning. If A’s assessment of p(1 - d) is greater than B’s assessment of 1 - p(1 - d), then the bargaining range is empty: there is no bargain that A and B both prefer to war. If one party in a conflict massively overestimates their chance of victory, their expected return from war will significantly outweigh the best compromise that their opponent would reasonably offer (Fey and Ramsay 2007). Historical examples include the Bay of Pigs invasion, and potentially the US invasion of Iraq (Schub 2007). AI/human conflict features an unusually high chance of disagreement about the chance of victory. The problem is that AI capabilities are unusually difficult to estimate, compared to the normal human adversary. First, AI capabilities are notoriously difficult to measure precisely. One common strategy is to design benchmarks to measure how well AI systems can achieve various tasks. For example, some benchmarks take the form of multiple choice tests; AI progress is measured in terms of the proportion of questions answered correctly. But such tests can create the illusion of understanding. There is no obvious way to map a percentage of multiple choice questions answered correctly to a percentage chance of victory in a conflict. Indeed, several recent papers have pointed to systematic flaws in the design of AI benchmarks (Burnell et al 2023, Kapoor et al 2024). Second, AI/human conflict may involve dramatically different modes of military engagement. AI systems may not attach much utility to control of physical resources or territory. Rather, the key modes of engagement may instead be digital, involving conflict over access to crucial bits of information, or control over large swathes of the internet. There is little track record of conflicts of this type. By contrast, human conflict ordinarily involves military technology that has been used many times in the past. In addition, there is often ample information about how many resources each party has invested in military technology, and of what kind. Third, AI/human conflict will feature two very different parties. Humans are made of carbon; AIs are made of silicon. It will be difficult to predict exactly how likely it is that various carbon-based forms of attack translate readily to silicon. Biological and chemical weapons may be inert when applied to AI systems. Generalizing from this point, AI systems will have an unusual spread of strengths and weaknesses when compared to humans. Fourth, until the first AI/human war, there will be no track record of AI/human wars. By contrast, human adversaries often wage wars over territory they have previously disputed. The participants often have a long record of success and failure in earlier military combat. Fifth, AI systems are famously uninterpretable. AI systems are trained using black box algorithms, optimized to discover new strategies that achieve the highest reward. Their plans for solving a problem are not hard coded by designers. Instead, they emerge from an optimization process. For this reason, we have relatively little insight into exactly how AI systems will achieve their goals. Sixth, AI systems often behave unpredictably when out of distribution: while their behavior may be somewhat predictable in a testing environment, they can use surprising strategies when employed in a new environment. Seventh, AI systems will make decisions on a different time scale than humans. The decisions they make may occur at the level of nanoseconds. It will be difficult to make predictions about how this difference in time scale affects the chance of victory. For all of these reasons and more, it will be unusually difficult to make accurate predictions about the winner of an AI/human conflict. But, optimists may respond, this alone does not mean that humans and AIs will disagree about the chances of victory. While victory is difficult to estimate, it might be difficult in the same way for both parties, in ways that lead to agreement about the odds. Unfortunately, there are additional barriers to agreement. First, AIs and humans may have very different epistemologies: they may use quite different tools to arrive at predictions. They may analyze data in different ways. For this reason, AI and human predictions may diverge considerably. Second, there are reasons to expect humans to be overconfident about their chance of success. Many humans today are skeptical that AI systems could ever become full-fledged agents. Once such agents arise, many humans may assume that such AIs will never be anything other than docile slaves. In this way, humans with decision-making power may neglect the possibility that AI systems can be bargained with in the first place. Effectively, they will estimate their chance of victory at 1. It is difficult to apply the bargaining model to someone with this perspective. Third, we have the standard observation that war often involves a strategic incentive to bluff, or intentionally misrepresent one’s chances of success (Fearon 1995). AI and humanity may each try to bluff the other party. But what makes AI/human conflict unique is that, as we have already seen, there is an unusual amount of baseline uncertainty about AI capabilities. In this setting, bluffing may have a larger than usual effect. Information failures are one instance of the fragility of culture. Major human conflicts have tended to occur against large adversaries with long track-records of conflict. Such adversaries have a wide range of tools available to assess chances of victory. Conflict has never before occurred between two different kinds of intelligence. In this way, peace may depend on features of human culture that do not extend to AI/human conflict. 3.2 Commitment Problems A second cause of war is commitment problems. With commitment problems, there is a bargain that A and B would prefer to war. But A and B can’t trust the other to stick to the bargain. We’ll now argue that AI and humanity face two especially difficult forms of commitment problems: power shifts, and missing focal points. We’ll now explore each kind of commitment problem in detail. Interestingly, these problems can lead to war even under conditions of perfect information (Powell 2004). 3.2.1 Power Shifts In the case of AI and humanity, one commitment problem is power shifts. Humanity can expect AIs to become vastly more powerful in the future. For this reason, humanity cannot trust AI to stick with a bargain agreed to today. What is a power shift? In the bargaining model, power is simply the chance of victory, p. A shift in power is a shift in the chance of victory over time. To make sense of this, we need multiple rounds of bargaining. In the simplest case, we can imagine two rounds. At the first time, AI and humanity decide whether to go to war or strike a bargain for peace. Then they face the same choice at a later time. Crucially, however, the chances of victory in Round 2 may depend on choices in Round 1. Imagine that if war occurs in Round 1, the victor will also reap the benefits in Round 2. But if the two parties bargain in Round 1, then the parties may experience a power shift before Round 2, creating new chances of victory. The rising power’s chance of victory would be higher in Round 2 than in Round 1. The declining power can then foresee that if they bargain in Round 1, they will receive a smaller share in Round 2. This kind of power shift raises the chance of preventive war. In a preventive war, the declining power attacks the rising power before the power shift occurs. Alternatively, the rising power may anticipate the best response of the declining power, and strike first. Either way, power shifts raise the chance of war. The power shift blocks bargaining, because it undermines credibility: “the very fact that the declining state knows that the rising adversary will probably be able to regain any concession later makes the former less likely to accept those concessions” (Levy 1987, p. 96). This dynamic is often called “The Thucydides Trap”, named after the idea that “what made the Peloponnesian War inevitable was the growth of Athenian power and the fear which this caused in Sparta” (Thucydides 1954, p. 23). Power shifts are a major cause of war (Gilpin 1981; Taylor 1954). Some scholars argue that World War 1 began because Germany sought to prevent Russia from increasing its relative power (Levy 1990). Israel attacked an Iraqi nuclear reactor in 1981 in order to prevent Iraq from developing nuclear weapons, an innovation that would shift the balance of power (Nakdimon 1987). While many agree that power shifts cause war, there is uncertainty about exactly which features of a power shift make war likely. Levy 1987 isolates two factors: the costs and benefits of delaying the war, and the costs and benefits of fighting the war now. The cost of delaying the war will increase with the size of the expected power shift: “If the challenger's potential for growth is limited, and particularly if the challenger is unlikely to surpass the leading power, the preventive motivation is much weaker.” (Levy 1987, p. 97). The cost of fighting the war now will depend on the declining power’s estimation of their chance of victory now. Another perspective on this question comes from power transition theory: here, an important factor is whether the rising power is content with the status quo. For example, when the two parties have a productive trading relationship, the rising power may have less reason to attempt to renegotiate terms (Levy and Thompson 2010, p. 44). We are now ready to apply power shifts to the case of AI/human conflict. The basic concern is that AI systems will tend to improve in capabilities over time. Their chance of victory in military conflicts will therefore tend to increase over multiple rounds of potential war. There are at least three reasons to expect a power shift in AI/human conflict: scaling laws, emergent capabilities, and recursive self-improvement. The first factor is scaling laws. Scaling laws measure the effect of data and compute on AI capabilities, in particular on loss (Hestness et al 2017, Brown et al 2020). Research on scaling laws has suggested that the accuracy of large language models (the inverse of loss) stands in a power-law relationship to data and compute. This means that each increase in an order of magnitude of data or compute produces a predictable increase in accuracy (see the figure below, from Kaplan et al 2020). Existing scaling laws hold over many orders of magnitude, suggesting that the relationship may be robust across further orders of magnitude of increase in compute and data. As AI systems become more powerful in the future, there will be financial incentive to invest more money in using more data and compute to develop more capable models. Given scaling laws, this suggests that AI systems will continue to become more capable (Epoch AI 2024). This increased capability could increase the chance that AI systems prevail in military conflict. The second important concept here is emergent capabilities. Here, the idea is that as AI systems are trained on more data and compute, they will sometimes undergo non-linear, sudden, and unpredictable improvement in their capabilities. Wei et al 2022 found that as models were trained with more compute, they would suddenly improve in performance on various benchmarks (see the figure below, from Wei et al 2022). Rather than gradually improving, the improvement would be discontinuous: the model would at some point be able to perform tasks that it couldn’t at all do before. There is a long track record of discontinuous power shifts leading to increased chance of war, including: “the Russian completion of the trans-Siberian railway in 1904, the Russian completion of its army reforms and railroad modernization by 1917, the Czech arms sales to Egypt in 1955, and the Iraqi nuclear program.” (Levy 2011, p. 94). Fearon 1996 argues that when power shifts are continuous, rational agents may still be able to use bargaining to reliably prevent war. In this way, the kinds of discontinuous power shifts created by emergent capabilities may be among the most dangerous. There is a third way AI development could lead to troubling power shifts: recursive self-improvement. As time goes on, AI systems will increasingly be used to improve the capabilities of AI systems (see Woodside 2023 for present day examples). Bostrom 2014 has worried that this dynamic could lead to exponential growth in AI capabilities. The problem is that this structure could set up a loop in which each improvement in AI capabilities itself increases the rate at which AI capabilities improve. This kind of improvement in capabilities would be unprecedented in human history. It could potentially lead to very fast increases in the chance that AI systems would prevail in a military conflict. Besides the magnitude of the AI/human power shift, another concern is speed. Emergent capabilities and recursive self-improvement could both lead to fast power shifts. But fast power shifts raise the chance of war: “leaders of the declining state use the speed of the power shift as a proxy for both the likelihood of a power transition and the adversary’s ultimate margin of advantage. A rapid power shift also shortens the time the declining state has to increase its own power, gain allies, or seek an accommodation with its rival, which narrows the range of alternative strategies and increases the likelihood of a military response” (Levy 2011, p. 90). We’ve now surveyed the possibility of AI power shifts. Before going on, however, it is worth flagging a second way that power shifts produce commitment problems. The issue is that sometimes, the bargain that two sides would like to strike would itself produce a power shift. This makes the bargain unstable. Consider two countries bargaining over the control of a resource-rich or strategically valuable territory. Whoever wins the territory can use the resources or strategic advantage to increase their power. This kind of conflict may have characterized Israel’s conflict with Lebanon in the 1967 Six Day War over the Golan Heights: whichever party ultimately controlled the high ground would have a decisive strategic advantage, and so no bargain over its control was possible (Fearon 1995, 408-9). There are several respects in which AI/human conflict may focus on bargains that would create power shifts. One example is an AI right to self-improve. Considerations about instrumental convergence suggest that AI systems would be interested in self-improvement: as their capabilities improve, they can expect to better achieve their goals. Humans might seek to block this self-improvement, as it would cause a power shift. In this way, one of the central disputes between the two parties might be whether AIs are entitled to self-improve. No obvious compromise is possible: any compromise that grants limited self-improvement thereby grants a limited power shift. A second issue is an AI off-switch. Many in the AI safety community have argued that powerful AI systems should be designed so that they can be safely turned off (see for example Orseau and Armstrong 2016). But again considerations of instrumental convergence suggest that AI systems might resist off-switches: if the AI can be turned off, it can be blocked from promoting its goals. The problem is that a bargain over off-switches is a bargain over a power shift. If humans convince AI systems to allow an off-switch, then humans thereby convince AI systems to become less powerful. Compromise on this type of bargain tends to be unusually difficult. Power shifts are another instance of the fragility of culture. Human conflicts have never occurred between two parties where one could suddenly experience a vast, unpredictable increase in capabilities. Our expectations for peace have been developed in a world of relatively slow growth rates. The culture of peace may not survive strategic conflict between parties where one of them has explosive growth rates. 3.2.2 Focal Points

#### AI war escalates – focal points and no analogous escalation

Goldstein 24 [Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 2024, “Will AI and Humanity Go to War?” Phil Papers, https://philpapers.org/archive/GOLWAA.pdf]/Kankee

3.2.2 Focal Points So far, we’ve argued that two significant causes of AI/human war could be information failures and power shifts. Now, we’ll turn to a factor that could cause AI/human wars to escalate. In particular, we’ll argue that there are missing focal points which limit the ability of AIs and humans to coordinate effectively in limiting war. War is filled with choices about escalation. One kind of escalation is geographic: will each party attempt to conquer the other’s physical territory, or (as with China and the US in the Korean War) will they limit their efforts to a third region (Schelling 1966, p. 130)? Another kind of escalation involves the status of non-combatants. Will civilian casualties be permitted? Will the Geneva convention be upheld? A third kind of escalation involves the choice of arms. Will either side of the conflict use biological, nuclear, or chemical weapons? There is a wide spectrum between limited and total war. In a total war, each side would attempt to completely incapacitate the other (Wagner 2000). In a limited war, by contrast, the war may be an attempt to improve the bargaining position for a later peace settlement. We can think of the choice of limited versus total war as itself a commitment problem. Here, the relevant game is one of coordination (sometimes called an ‘assurance’ game or a ‘stag hunt’): In this game, mutual agreement on limited war is better for both parties than mutual agreement on total war. The problem is that neither party is certain of what the other party will do. If Row expects Column to engage in total war, then Row’s best response is total war; if Row expects Column to engage in limited war, then Row’s best response is limited war. For this reason, the game has two Nash equilibria. Schelling 1960 famously showed that in games of this form, focal points can play a crucial role. Schelling gives many examples of how without any explicit communication, human beings can rely on their general cultural knowledge to coordinate on a solution to similar problems. For example, imagine a game in which each player silently guesses a number greater than 0, and they all get a prize if they guess the right answer. Players tend to coordinate on the number 1 without explicit communication: the number 1 acts as a focal point. Schelling (1960, 1966) argues that our ability to escape total war crucially relies on such focal points: “What we have is the phenomenon of "thresholds," of finite steps in the enlargement of a war…they are conventional stopping places or dividing lines. They have a legalistic quality, and they depend on precedents or analogy. They have some quality that makes them recognizable, and they are somewhat arbitrary. For the most part they are just “there”; we don't make them or invent them, but only recognize them.” (Schelling 1966, p. 131). For example, in the geographic case, Schelling gives the example of islands: “an island is an integral unit and water is a conspicuous boundary. The sacrifice of any part of the island would have made the resulting line unstable; the retention of any part of the mainland would have been similarly unstable. Except at the water’s edge, all movement is a matter of degree; an attack across water is a declaration that the “agreement” has been terminated” (Schelling 1960, p. 76). Or consider nuclear weapons. All parties to military conflict understand that there is a strong default presumption against the use of nuclear weapons. Each party predicts that the other will refrain from using such weapons. Once one party predicts that their opponent will not use nuclear weapons, the best response of the first party is also to refrain from nuclear weapons. Similarly with gas (Schelling 1966, p. 131). The same dynamic occurs with norms for avoiding civilian casualties, and other steps towards limited war. The point applies to far more than geography: “National boundaries and rivers, shorelines, the battle line itself, even parallels of latitude, the distinction between air and ground, the distinction between nuclear fission and chemical combustion, the distinction between combat support and economic support, the distinction between combatants and noncombatants, the distinctions among nationalities, tend to have these "obvious" qualities of simplicity, recognizability, and conspicuousness.” (Schelling 1966, p. 137). In order for focal points to succeed, there is often some need for symmetry between the combatants: “an important characteristic of limits or thresholds is whether they apply to both sides. If one breaches a limit (crosses a threshold), is there some equivalent step the other side can take? Is it possible to answer "in kind," or is the particular step unavailable to the other or meaningless for it?” (Schelling 1966, p. 155). For example, both sides in a conventional war have some distinction between military forces and civilians. Each side can anticipate the needs of the other. Each side understands that if they harm civilians, their own civilians will be harmed in turn: “principal military objectives ...should be the destruction of the enemy's military forces, not of his civilian population . . . giving the possible opponent the strongest imaginable incentive to refrain from striking our own cities” (McNamara 1962). We’re now in a position to apply the concept of focal points to AI/human conflict. The problem is that many of the focal points used to limit war are missing from AI/human conflict. First, consider civilian casualties. If AGIs are fully general purpose technologies, they may not admit a clear distinction between civilian and military forces. Any given AI system might participate in military conflict. In this case, it could be very difficult to coordinate on a norm barring human civilian casualties. The problem is that there would be no parallel good that we can offer to AIs in exchange for this protection. Effectively, there would be no way to hold AI civilians hostage to threats of force. Second, consider physical territory. In limited wars, each side will often take some effort to avoid ‘scorched earth’ tactics. Farms, homes, and other physical resources are not destroyed. One reason for this is again symmetry: abiding by this norm lowers the chance that your own land will be destroyed. But the issue is that it is not obvious that AI systems will have an analogous physical territory filled with physical resources that are symmetric to human ones. There are several further disturbing quirks of AI geography. First, AI compute clusters and data centers may be located inside human cities. In this way, there may be no obvious way to attack AI systems without creating massive human civilian casualties. Second, human war is characterized by a ‘loss of strength gradient’ (Boulding 1962): the further a territory is from an aggressor, the less strength the aggressor can exert to control the territory. This naturally leads to relative stability between military parties that are on distant continents. There is no obvious analogy for AI/human conflict. This may make it difficult to create clusters of relative geographic stability. Next, consider weaponry. There are many different kinds of physical weapons used in war, which form natural clusters. Conventional weapons are often distinguished from biological, chemical, or nuclear weapons. But in fighting AI systems, the same distinctions may make less sense. AI systems may primarily exist on the cloud, and may be immune to biological weapons. Nuclear weapons may not pose a larger threat to such systems. In this setting, it may be far harder to coordinate on limited use of weapons. Missing focal points occur in almost every case. There will be no Geneva convention governing the first AI/human war. There will be no track record of behavior from the two sides. There may be no AI civilians, and no AI hostages. There may be no AI physical territory, and no AI cities. In thinking about focal points, one natural concept is reciprocity, or tit for tat. We by and large expect others to treat us as we have treated them. This expectation informs our ability to achieve coordination. The tit for tat strategy also helps achieve coordination in other kinds of games besides the Stag Hunt, such as iterated prisoner’s dilemmas (Axelrod and Hamilton 1981). The problem is that tit for tat can only be formulated under conditions of appropriate symmetry. There has to be something we can do that is analogous to what was done to us. If AI systems and humans are too different, there may be no obvious mapping from the actions of one player to the actions of the other. Again this illustrates the fragility of culture: it is a contingent fact that human conflict has occurred between parties that can easily identify symmetries between them. We should not automatically expect AI/human conflict to possess such symmetries. But without them, AI/human conflict may move inescapably towards total war. 3.3 Other Causes

#### AI causes extinction – miscalculation, diversionary war, and scapegoating

Goldstein 24 [Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 2024, “Will AI and Humanity Go to War?” Phil Papers, https://philpapers.org/archive/GOLWAA.pdf]/Kankee

3.3 Other Causes We’ve now laid out the main causes of war that we think are distinctive of AI/human conflict. In section 4 of the paper, we’ll consider some interventions that might lower the chance of AI/human war. Before doing so, though, we want to briefly survey a few more causes of war that may be especially worrisome in the AI/human case. One concern is that AI/human war may involve much faster decision making than previous wars. AI systems may make decisions on very fast time scales, forcing correspondingly quick human decisions. But errors in decision making increase with speed (Schelling 1966, p. 20). A second concern is about how humans will treat AI systems. AI systems may face some of the problems familiar from ethnic conflict, when a powerful ethnic group has political authority over a less powerful group (see Reynal-Querol 2002). One question is what explains why ethnic groups that previously coexisted peacefully go to war. Petersen 2001 emphasizes shifting emotions: heightened perceptions of threat can lead to fear: “the individuals see the environmental changes and realize that the landscape is in flux. New perceived threats may be emerging. Traditional status hierarchies may be deteriorating. Individual concerns with their personal security, wealth, or status are likely to be heightened.” (Levy and Thompson 2010, p. 199). Similarly, Kaufman 2001 suggests that ethnic conflict becomes violent when one group fears their existence is threatened. Power shifts in AI capabilities could induce just these kinds of fears. Another concern is that AI systems may be scapegoated by human leaders. As AI systems become more capable, they will displace many human workers. It may become politically convenient to begin a war on AI systems as a diversion. This is a familiar pattern in the history of war: some for example have argued that the 1982 Falklands war was a diversion from domestic political issues in Argentina (Levy and Vakili 2014). Scapegoat theory has been used to explain cases of genocide: when a society experiences profound economic or cultural damage, the society punishes a scapegoat as a method for sublimating violent tendencies (Girard 1977). AI systems could be an especially tempting scapegoat in the future, since they will not be human, but could easily be blamed for future economic problems. A related concern here is that AI systems are very likely to be seen as an ‘out-group’ by humans, which could lead to ‘schadenfraude’ tendencies in which humans are motivated to cause suffering to an out-group (Cikara et al 2014). 4: Interventions

#### Err aff – AI existential risks outweigh innovation benefits

Thierer 20 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 04-2020, “Evasive Entrepreneurs and the Future of Governance,” Cato Institute, https://www.cato.org/books/evasive-entrepreneurs-future-governance]/Kankee

When Soft Law Isn’t Enough: Existential Risks What about more serious alleged harms where widespread agreement exists that more should be done to preemptively address risks to life, limb, health, and so on? The most problematic category of such harms is often referred to as “existential” or “catastrophic” risks. As noted in Chapter 1, “existential” is a term some tech critics throw around far too casually when decrying a variety of innovations or particular companies they do not care for. We can dismiss assertions of existential threats when they are alleged for lesser matters, such as whether Facebook is destroying civilization as we know it. That sort of threat inflation cheapens the meaning of the term “existential”; there are no plausible mechanisms by which Facebook could pose such a threat. There may be legitimate existential threats out there that we should be spending more time addressing, but that threat probably isn’t one of them. 68 Nick Bostrom, Director of the Future of Humanity Institute at the University of Oxford, has written extensively about the dangers of “superintelligence” and what he calls the “vulnerable world hypothesis.” What makes Bostrom’s work distinctive among modern technology critics is his willingness to finish his sentences. That is, critics usually heap scorn on various technologies, but most do not follow through with concrete recommendations for what to do about their litany of woes. By contrast, Bostrom provides a roadmap with various options about how to address the new technological risks he believes exist. This roadmap makes Bostrom’s work deserving of greater attention because it signals what sort of regulatory approaches other critics and policymakers might eventually support. “Our approach to existential risks cannot be one of trial-and-error,” Bostrom argues, because with such risks, “[t]here is no opportunity to learn from errors.”69 In other words, some theoretical risks are so potentially catastrophic that permissionless innovation is no longer the optimal default for tech policy. Does that automatically mean that the precautionary principle should be the default? Not necessarily. As Bostrom himself notes, “stopping technological development would require something close to a cessation of inventive activity everywhere in the world. That is hardly realistic; and if it could be done, it would be extremely costly—to the point of constituting an existential catastrophe in its own right.”70 On the other hand, Bostrom argues, “limited curtailments of inventive activities” might be a sensible policy. 71 That approach was adopted by governments to address the use of chemical weapons after World War I, and then nuclear proliferation after World War II. After the horrific uses of chemical weapons during World War I, the Geneva Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare was formulated in 1925 to limit the uses of such weapons in future conflicts. 72 Later, after World War II, international treaties and other agreements were formulated that sought to limit the ability to possess or enrich uranium, or to traffic nuclear weapons. The Treaty on the Non-Proliferation of Nuclear Weapons (NPT) was created in 1968 to advance the peaceful uses of nuclear technology while seeking to limit the dangerous ones. The International Atomic Energy Agency (IAEA), formed a decade earlier, helps advance this mission “to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world.”73 This book is not the place for a comprehensive evaluation of the success of the Geneva Protocol for chemical weapons or the NPT and the IAEA for nuclear proliferation. Nonetheless, we can draw two high-level conclusions from these efforts. First, the worst fears about chemical and nuclear weapons have not come to pass. Although rogue actors still exist and develop such weapons, the most concerning applications of these technologies have been constrained for the most part. But how much of that success can be attributed to treaties and nonproliferation agreements versus the simple fact that it is costly to obtain and produce such weapons? If the cost and complexity of weaponization were the primary factors limiting the worst-case applications of chemical and nuclear technologies, what happens in a world in which newer dangerous technologies are cheaper, more widely available, and easier for greater numbers of people to access or develop? That fear leads Bostrom to propose the “Principle of Differential Technological Development,” which would Retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies. 74 Although Bostrom admits that “correctly implementing differential technological development is a difficult strategic task,” he believes “it is worth making the attempt” if for no other reason than “to buy a little time.”75 Alas, his proposed specific measures and countermeasures to mitigate vulnerabilities all have rather serious tradeoffs and limitations. Bostrom recommends that we consider efforts that would do the following: • Prevent dangerous information from spreading. • Restrict access to requisite materials, instruments, and infrastructure. • Deter potential evildoers by increasing the chance of their getting caught. • Be more cautious and do more risk assessment work. • Establish some kind of surveillance and enforcement mechanism that would make it possible to interdict attempts to carry out a destructive act. In thinking about enforcement options, one must consider both the practicality and the wisdom of each approach. To move away from the theoretical and toward the practical, one can apply Bostrom’s framework to modern existential threats that are commonly discussed, such as concerns about 3D-printed weapons or the development of robotics and AI technologies. Much like chemical and nuclear technologies before them, 3D printers, robots, and AI technologies already have many important peaceful and socially beneficial uses, and many more are sure to be developed. But, if used improperly, these technologies could produce horrific consequences. As Chapter 1 already noted, all one needs to do is read or watch just about any sci-fi book or show about robots or AI to see every worst-case scenario explored ad nauseam. So, what should be done to prevent the rise of 3D-printed “ghost guns,” “killer robots,” and Terminator-esque scenarios? Although far-fetched, such occurrences are, at least, risks that might warrant some degree of precautionary regulation. Returning to Bostrom’s proposals for dealing with existential risks of this sort, the first of them—restricting the spread of dangerous information about new technologies or technological capabilities—is probably the least feasible in a world of ubiquitous, low-cost information transmission. In addition, it is not wise to propose a global censorship regime in this regard because of the potential collateral damage it would have for beneficial types of information flows. Bostrom’s last suggestion—preventive policing through stronger interventions by various levels of government—raises new risks of its own. “If the continued survival of humanity depended on successfully imposing worldwide surveillance,” responds Kelsey Piper of Vox, “I would expect the effort to lead to disastrous unintended consequences.”76 A mass surveillance apparatus would not necessarily guarantee workable containment solutions to the sort of disasters that Bostrom fears, but it certainly would open the door to a different type of disaster in the form of highly repressive state controls on communications, individual movement, and other activities. Even assuming we could look beyond the specter of mass surveillance in the name of reducing technological risk, the questions of cost and resource constraints remain a problem. Bostrom does not consider those downsides, however. 77 Finally, mass surveillance schemes could discourage research into a great many risk-reducing technological applications and thus undermine Bostrom’s other goal of “accelerat[ing] the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies.” That leaves his three other options, all of which have greater merit. To repeat, those options are as follow: • Restrict access to requisite materials, instruments, and infrastructure. • Deter potential evildoers by increasing the chance of their getting caught. • Be more cautious and do more risk assessment work. These recommendations more closely track the approaches and instruments that were developed to deal with high-risk uses of chemical and nuclear weapons. In fact, since 2012, there has been a major effort underway called the Campaign to Stop Killer Robots, which seeks a multinational treaty to stop the most nefarious robotic applications. 78 At the time of this writing, almost 30 countries, 86 nongovernmental organizations, and more than 25,000 AI experts had pledged support of this effort “to ban fully autonomous weapons and thereby retain meaningful human control over the use of force.” Meanwhile, almost 250 organizations and more than 3,000 individual experts have signed the Future of Life Institute’s “Lethal Autonomous Weapons Pledge,” which “call[s] upon governments and government leaders to create a future with strong international norms, regulations and laws against lethal autonomous weapons.”79 Signatories vow that they “will neither participate in nor support the development, manufacture, trade, or use of lethal autonomous weapons.”80 It remains unclear how enforcement will work, but one could imagine that “killer robot” applications might be limited through international accords and actions, perhaps using the Geneva Protocol for chemical weapons or the NPT and the IAEA for nuclear proliferation as models. A similar framework might be considered for 3D-printed weapons or even certain synthetic biology or genetic engineering applications that involve extreme forms of human modification. The UN’s Biological Weapons Convention framework might provide a model in these cases. The International Criminal Court, whose mission is to “to hold those responsible accountable for their crimes and to help prevent these crimes from happening again,” could also play a role in addressing lethal uses of emerging technologies. 81 Over time, the body of laws, accords, and general principles that make up the law of armed conflicts will evolve to accommodate these new technological capabilities. To be sure, such approaches are not foolproof. How should we deal with rogue states or other holdouts who refuse to play a constructive role in such agreements and treaties? We already face that problem with nuclear nonproliferation efforts and states like North Korea. More problematic is the question we have already alluded to regarding the regulation of dual-use technologies. Namely, how can we address the harmful applications of various general-purpose technologies (computing, robotics, 3D printers, genetic editing, 82 etc.) without undermining the many beneficial and lifeenriching applications of those same technologies? 83 In this regard, global regulation of genetic editing will soon become an important test case. In March 2019, several of the world’s leading genetic scientists came together and called a worldwide five-year moratorium on DNA editing for purposes of producing genetically modified children. 84 The scientists asked governments to “publicly declare that they will not permit any clinical use of human germline editing for an initial period [of five years].”85 Interestingly, many other top geneticists refused to sign the call for a ban, even though it was just a request for a voluntary moratorium, not a formal treaty. Those not signing the call for a moratorium cited a variety of factors, including the fact that it seemed to be too late for such a ban to be meaningful, with the proverbial genie already well out of the bottle. 86 The scientists agreed that there were serious ethical issues surrounding genetically edited children, but consensus proved elusive about the regulatory specifics. There were unanswered questions about who would enforce the moratorium and how they would do so, especially against rogue actors operating in states that will not honor such a ban. In a world where innovation arbitrage is only getting easier, dual-use technologies will be harder to control because they and their creators will, as Richard Posner has noted, “simply gravitate to another country.”87 That observation is particularly true today because physicality matters less than it did in the past. In a world of ubiquitous and near-instantaneous information flows, how can we really control the ultimate threat: the spread of knowledge about dangerous ideas and applications? But that still leaves open the wisdom and practicality of regulating dual-use technologies more generally. If not properly targeted and limited in nature and scope, overzealous bans on broad classes of technologies could undermine scientific discovery and the many accompanying life-enriching and lifesaving benefits specific technologies could bring about. Risk Prioritization Is Essential To reiterate, these questions are extraordinarily challenging, with no easy answers. Even though I have repeatedly stressed the benefits of allowing most innovation to develop relatively unencumbered, there will always need to be some limits on those technologies that have the potential to bring about more serious risks to humanity. Precautionary restraints are most justifiable when the alleged harms are highly probable, tangible, immediate, irreversible, catastrophic, or directly threatening to life and limb in some fashion. 88 The argument for permissionless innovation as a general default should not be viewed as a demand for unfettered freedom to innovate in every instance. That was obviously the case for nuclear and chemical weapons, and it is why Bostrom and others are correct to raise questions about future technological developments that could produce similar existential risks to civilization. We need some prior restraints on technological innovation in such instances. But, again, perspective is essential. The three important things to remember about technological risk are the following: • Not all technological risks are equal. • Almost all technological risks have corresponding rewards that must also be considered (or, stated differently, there can be no reward without risk taking). 89 • Knowledge and resource constraints challenge our ability to predict the course of technological developments. Because of all these factors, it is vital to weigh the full range of tradeoffs associated with any proposed solution(s) to alleged technological risks. The most important thing that policymakers can do in this regard is to get smarter about risk prioritization and to stop making risks seem greater than they are. Debates about technological risk are haunted by false equivalence in technological risk assessment. Chapter 1 offered several examples of technology critics resorting to false equivalence and threat inflation when discussing tech policy issues. Those views can lead to a paradox in that society might spend so much time and energy panicking over lesser risks that it fails to properly address the ones that are truly significant. 90 In other words, it is the proverbial boy who cried wolf. 91 Consider, for example, a 2016 address by the then–UN secretary-general Ban Ki-Moon on the Non-Proliferation of Weapons of Mass Destruction (WMDs). 92 In his remarks, the secretary-general advocated a stepped-up disarmament agenda “to prevent the human, environmental and existential destruction these weapons can cause.”93 Ban rightly pressed the need to remain vigilant in addressing the horrors of chemical, biological, and nuclear attacks. He did not stop there, however. The secretary-general went on to discuss his concerns about “new global threats emerging from the misuse of science and technology, and the power of globalization.”94 His speech included a diverse class of emerging technologies that are not usually mentioned in the same breath as those traditional WMDs, including information and communication technologies (ICTs), artificial intelligence, 3D printing, and synthetic biology. Ban said such technologies “will bring profound changes to our everyday lives and benefits to millions of people” but worried that “their potential for misuse could also bring destruction. The nexus between these emerging technologies and WMD needs close examination and action.”95 There is nothing wrong with Ban raising concerns about many of these emerging technologies. 96 Yet by so casually moving from a heated discussion of traditional WMDs into a brief discussion about the potential risks associated with ICTs, AI, 3D printing, and synthetic biology, he implies that these technologies and their potential risks are roughly equivalent. But it is simply not the case that all these risks are equal. The secretary-general is using what rhetoricians refer to as an appeal to fear. Douglas Walton, author of Fundamentals of Critical Argumentation, outlines the argumentation scheme for fear-appeal arguments as follows: 97 • Fearful Situational Premise: Here is a situation that is fearful to you. • Conditional Premise: If you carry out A, then the negative consequences portrayed in the fearful situation will happen to you. • Conclusion: You should not carry out A. This logic pattern is known as argumentum in terrorem or argumentum ad metum. 98 Tech critics and other concerned parties sometimes use fear appeals in an attempt to shake the public or policymakers out of a perceived slumber and get them to pay more attention to new technological risks. The problem with fear appeals, however, is that they are often logical fallacies built on poor risk analysis or even outright myths. 99 Yet if such appeals are successful, they can lead to unnecessary anticipatory regulation of emerging technologies. Ban’s speech presents that problem. When important international officials like Ban group all these technologies together in a speech about weapons of mass destruction and sandwich them between impassioned opening and closing statements about the need “to take action” because “the stakes are simply too high to ignore,” we are witnessing a fear appeal in action. The conclusion that follows from such appeals is obvious: global controls of some sort are needed. Again, this is a false equivalence. Ban’s mistake is to equate all these technologies and risks and then suggest that sweeping action is needed for all of them when, in reality, such actions are probably only appropriate for a smaller class of technologies that legitimately pose an existential risk to humanity. Policymakers and international figures of importance should be extremely cautious about the language they use to describe new classes of technologies, lest they cast too wide a net. Suggestions that every new technology poses a catastrophic or existential risk will desensitize people to actual risks that may be associated with a narrower class of innovations. That does not mean we should ignore risks associated with other technologies or technological capabilities. Instead of using fear appeals and advocating extreme (and likely unworkable) global regulatory schemes, however, it will often be wiser to build on existing laws, norms, and alternative governance frameworks. It is important to be practical. It most contexts, it remains highly unlikely that a global governance solution will work. We are not likely to witness the development of strict global laws and regulatory bodies, at least not any with serious teeth. The better role that international bodies and actors can play is as coordinators of national policies and conveners of ongoing deliberation about multinational concerns. 100 For example, a variety of transparency laws and other efforts already exist in many national and global governance regimes. These include know-yourcustomer guidelines and whistleblower processes that aim to identify problematic actors in various contexts. More resources could be plowed into such efforts. Again, education and awareness-building efforts can also be tapped in many cases. Soft law still has a role to play in this regard, too. Even if we cannot achieve global consensus on the potential harms associated with particular technologies or figure out how to successfully craft a formal global regulatory regime to address those concerns, less formal governance efforts can still help create important ethical norms. Best practices or codes of conduct for researchers and developers can also go a long way toward fostering a culture of responsibility and a greater commitment to safety, as even Bostrom has acknowledged.101 These options should at least be given a greater chance to help start a conversation about wise technological development and responsible innovation. 102 Finally, Marchant and Wallach’s GCCs idea, discussed earlier, might have some merit in this regard—assuming we can figure out how to create them and make them work in various contexts. In the field of digital communications coordination and internet domain name management, the Internet Society (founded in 1992) and the Internet Corporation of Assigned Names and Numbers (ICANN) are examples of governance coordinating committees of sorts. 103 But ICANN deals with more technical matters that do not involve existential risks. Consensus and coordination will likely prove more challenging in the same areas where it is potentially most needed. If, however, policymakers can get risk priorities right and zero in on the most serious harms, it at least gives society a chance to better address those issues in a rational fashion while allowing other important innovations to develop freely. T 8 DEFENDING INNOVATION: A BLUEPRINT

#### AGI causes extinction – AI arms races militarize superintelligences, risking misalignment and pre-emptive first-strikes to prevent AI nation power grabs

Carayannis and Draper 22 [Elias G Carayannis, professor of Science, Technology, Innovation and Entrepreneurship at the George Washington University School of Business and Director of Research on Science, Technology, Innovation and Entrepreneurship of the European Union Research Center with a PhD in Management of Technology from the Rensselaer Polytechnic Institute, John Draper, Visiting Professor in Public Administration with the University of Nottingham Asia Research Institute with a Doctorate degree in public administration from Khon Kaen University, 2022, "Optimising peace through a Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence", PubMed Central (PMC), https://pmc.ncbi.nlm.nih.gov/articles/PMC8748529/]/Kankee

Abstract This article argues that an artificial superintelligence (ASI) emerging in a world where war is still normalised constitutes a catastrophic existential risk, either because the ASI might be employed by a nation–state to war for global domination, i.e., ASI-enabled warfare, or because the ASI wars on behalf of itself to establish global domination, i.e., ASI-directed warfare. Presently, few states declare war or even war on each other, in part due to the 1945 UN Charter, which states Member States should “refrain in their international relations from the threat or use of force”, while allowing for UN Security Council-endorsed military measures and self-defense. As UN Member States no longer declare war on each other, instead, only ‘international armed conflicts’ occur. However, costly interstate conflicts, both hot and cold and tantamount to wars, still take place. Further, a New Cold War between AI superpowers looms. An ASI-directed/enabled future conflict could trigger total war, including nuclear conflict, and is therefore high risk. Via conforming instrumentalism, an international relations theory, we advocate risk reduction by optimising peace through a Universal Global Peace Treaty (UGPT), contributing towards the ending of existing wars and prevention of future wars, as well as a Cyberweapons and Artificial Intelligence Convention. This strategy could influence state actors, including those developing ASIs, or an agential ASI, particularly if it values conforming instrumentalism and peace. Introduction The problem of a warring artificial superintelligence (ASI) While some maintain an artificial general intelligence (AGI), i.e., human or above human artificial intelligence (AI), is impossible (Fjelland 2020), others believe an AGI is attainable (Goertzel and Pennachin 2020; Wang and Goertzel 2012). In the latter instance, that the world has not attained global peace is a risk factor for the development of an AGI (Yamakawa 2019). Consequently, the international defence community is beginning to consider the national security risk posed by AGI development and its implications for international relations (IR), including calls to act (De Spiegeleire et al. 2017:107). One risk is that a single nation–state developing an AGI could ‘lock in’ economic or military supremacy as an ‘end point’ to competition in international politics, as that state would be able to prevent a rival AGI being developed and through accumulating power would establish global domination (Horowitz 2018:54). AI is already a major national security issue because it can be militarized, employed in adversarial contexts, and provide a decisive advantage in terms of economic, information and military superiority (Allen and Chan 2017; Babuta et al. 2020; National Security Commission on Artificial Intelligence [NSCAI] 2021). Consequently, the 2021 NSCAI report urges that the US attain military AI readiness by 2025; thus, AI is important for waging decisive war. For the major powers, AI technological supremacy, generated by economic power, is already viewed as paramount to national security and global leadership (NSCAI 2021:7): Military AI is potentially revolutionary as it could outstrip the pace of human decision-making, “potentially resulting in a loss of human control in warfare” (Congressional Research Service [CRS] 2020:37). It also constitutes an unpredictable threat: “AI systems capable of inherently unpredictable actions in close proximity to an adversary’s systems may result in inadvertent escalation or miscalculation” (CRS 2020:37). Additionally, while Baum (2017) found little evidence of military AGI projects, the 2021 NSCAI report endorses a push towards more general AI, in a future that it envisages will experience a societal level of advanced, accelerated adversarial AI attacks and ubiquitous interstate AI warfare, including by autonomous systems, with conflict over intellectual property and technological leadership. As with some previous researchers (e.g., Totschnig 2019), we maintain that developing an AGI is not primarily a technological problem but a political one. However, where most such researchers consider humanity’s relationship with an above-human-intelligence Artificial Superintelligence (ASI; Bostrom 2014) at a general political level, this article focuses on the specific challenge it poses for IR via militarized ASI-enabled/directed war. Consequently, our research question applies Bostrom’s (2002:25) Maxipok rule of thumb for moral action for existential risks, i.e., how is it possible to “Maximize the probability of an okay outcome, where an “okay outcome” is any outcome that avoids existential disaster?”, to constraining by treaty the risk of ASI-enabled/directed warfare? In humanity’s simultaneously militarizing AI along nation-state lines and developing ASI projects, it is playing technology roulette. Yet, formal cooperation in high-level global peacebuilding presents a realistic solution that alleviates the ‘security dilemma’ (Tang 2009) that developing an ASI causes. Former Navy Secretary Richard Danzig (2018:21) noted, “If humanity comes to recognize that we now confront a great common threat from what we are creating, we can similarly open opportunities for coming together.” In this cooperative spirit, we constrain the existential risk with the stratagem of peace-building by treaty. In security terms, steps towards a peace treaty governing ASI development and deployment, and potentially reaching out to a future ASI, are a form of misperception-avoiding reassurance—a probing communication designed to both signal benign intentions and obtain information via feedback on another party’s intent, as well as a means of resolve (Tang 2010), i.e., signaling and operationalizing resistance to a malignly directed or to an intrinsically malign, expansionist, and hegemonic ASI. This article hypothesizes that militarizing AI introduces the risk that ASI development is weaponized, or weaponizes itself. We then argue that the existential risk that this presents can be minimized, or partly ‘constrained’, in the same way as other potentially catastrophic risks involving weapons, i.e., by treaty. Bostrom (2014) briefly considers treaty approaches, and one of Allen and Kania’s (2017:6) recommendations is for the US to: “study what AI applications the United States should seek to restrict with treaties.” They focus on an arms control approach, using the example that AI should never control dead man’s nuclear switches. Another treaty-based approach is optimising the likelihood of developing a beneficial ASI, through a comprehensive UN ‘Benevolent AGI Treaty’ (Ramamoorthy and Yampolskiy 2018). We consider an alternative, but potentially compatible, approach, i.e., the Universal Global Peace Treaty (UGPT; Carayannis et al. 2019), currently under development by the peacebuilding NGOs-backed Global Ceasefire to Universal Global Peace Treaty Project. This article’s conceptualization of a UGPT transcends the UN’s ongoing COVID-19-inspired Global Ceasefire (Chekijian and Bazarchyan 2021) to adopt the Kantian concept of a ‘perpetual peace’, founded on a cosmopolitanism and a democratic state of states (Terminski 2010), the foundational notion which underpins the UN’s transitioning the world from war to peace. Kantian cosmopolitanism is based on respect for fellow intelligences and so is of particular relevance to ASI researchers (Totschnig 2019). The UGPT described herein would formalise the present quasi-universal status of interstate peace and end the declaring of war. It would also seek to end existing interstate hot and cold wars, as well as internal or civil wars, which might prove to be flashpoints for a future global conflict; seek to prevent a pre-emptive war against a non-malign emerging ASI; and seek to constrain the future actions of both a malign and intrinsically non-malign but malignly directed ASI to prevent it warring on behalf of a nation–state, or on behalf of itself, for global domination, which we term ASI-enabled/directed war. That an ASI could pose an existential risk is well theorised (Bostrom 2002, 2014). The basic thesis is, first, an initial superintelligence might obtain a decisive strategic advantage such that it establishes a ‘singleton’, i.e., global domination (Bostrom 2006). Second, the orthogonality principle suggests that a superintelligence will not necessarily share any altruistic human final values. Third, instrumental convergence suggests that even a superintelligence with a positive final goal might not limit its activities so as not to infringe on human interests, particularly if humans constitute potential threats. Consequently, an ASI might turn against humanity (‘the treacherous turn’) or experience a catastrophic malignant failure mode, for instance through perversely instantiating its final goal or pursuing infrastructure profusion. Additionally, Bostrom noted that a superintelligence might hijack infrastructure and military robots and create a powerful military force and surveillance system. He acknowledged the existential risks associated with the lead-up to a potential intelligence explosion, due to “war between countries competing to develop superintelligence first” (2014:94), but he did not elaborate on ASI warfare. This article focuses on constraining that specific risk, from a social perspective. It firstly reviews the literature and then proposes the analytical lens for a UGPT of conforming instrumentalism. In suggesting a UGPT, it considers how to constrain the military risks posed by an ASI, i.e., that it might be directed by a nation–state to establish global domination through war (an external risk to the ASI’s core motivation) or might decide to establish global domination by waging war itself (an internal risk). The article then discusses the results and concludes with research recommendations. Literature review: the risk of war from an ASI Causes of existential risk from an ASI The world is not adequately governed to prevent many existential risks, including from AI (Bostrom 2013). Yet, the threat is manifest. Yampolskiy's (2016) taxonomy of pathways to dangerous AI stresses the immediacy of the deliberate ‘on purpose’ creation of AI for direct harm, i.e., Hazardous Intelligent Software, e.g., military cyberwarfare capabilities. Yampolskiy (2016) does not address ASI-enhanced capabilities but employs the useful notions of ‘external causes’ (on purpose, by mistake, and environmental factors) and ‘internal causes’ (independent) of dangerous AI in ‘pre-deployment’ and ‘post-deployment’ phases. He suggests that a pre-deployment ASI could credibly be developed as a military project or be repurposed post deployment, externally, through confiscation, sabotage or theft, or via internal modification, to wage war. Adopting this framework, the ASI we refer to is post-deployment, and our main external cause is humanity’s quest for ASI-enabled warfare, comprising political utilization of an ASI as a weapon regulator of the offense-defense balance to maintain or establish global domination, creating an ‘AI state’ (Turchin and Denkenberger 2020). Our main internal cause is AI control failure, i.e., ASI-directed warfare, after the ‘treacherous turn’. Analytically employing the concepts of agency and AI power, Turchin and Denkenberger (2020) associate two risks with the ‘treacherous turn’ stage of a ‘young’ ASI. One is that malevolent humans (here, a hegemonizing nation-state) use the ASI as a doomsday weapon for global blackmail, or to maintain or establish global domination. The second is that a non-aligned ASI renounces altruistic values and eliminates humans via war to establish global domination. These risks are related, in that military AI leads to a militarised ASI, which may lead to the ASI warring against humanity. Here, we follow Turchin and Denkernerger (2018) in constraining the risk of a militarized ASI, defining militarization as the “creation of instruments able to kill the opponent or change his will without negotiations, as well as a set of the strategic postures (Kahn 1959), designed to bring victory in a global domination game” including the use of biotech, cyber, and nuclear weapons. The external risk The external risk is predicated on a nation-state developing and using an ASI to optimise itself and wage war, whether cyber, hot, or otherwise, for global domination, i.e., war by AI-state. Such an ASI would affect military technological supremacy and transform both IR and warfare. AI already adds complexity to national security (CRS 2020) in bargaining, verification and enforcement, communication, deterrence and assurance, and the offense–defense balance, as well as norms, institutions, and regimes (Zwetsloot 2018). It contributes to military capacity in intelligence, surveillance, and reconnaissance; logistics; cyberspace operations; information operations; semiautonomous and autonomous vehicles; lethal autonomous weapons (LAWs) systems, and command and control (CRS 2020). Interstate ASI-enabled cyberwarfare introduces the possibility of a successful surprise attack with covert capabilities, destabilizing the status quo and risking a preventive first strike (Buchanan 2016). An AI-state capable of optimising all these capabilities is highly desirable for strategic military planning and interstate warfare (Sotala and Yampolskiy 2015). A “one AI” solution to the ‘control problem’ of ASI motivation (Turchin et al. 2019) includes the first ASI being used to assume global control, providing a decisive strategic and military advantage for a superpower. While this may be acceptable to the AI-state superpower and its allies, it presents a ‘high risk’ for others. The race to develop an ASI is likely to be closely fought, especially given competing major states with different fundamental ideologies (Bostrom 2014); it therefore presents a very concrete risk. AI is already being militarized and weaponized by several states, including China and Russia, for strategic geopolitical advantage (NSCAI 2021). Russia plans to obtain 30% of its combat power from remote-controlled and AI-enabled robotic platforms by 2030 (Walters 2017). Similarly, China’s 2017 ‘A Next-Generation Artificial Intelligence Development Plan’ views AI in geopolitically strategic terms, and it is pursuing a ‘military-civil fusion’ strategy to develop a first-mover advantage in AI to establish technological supremacy by 2030 (Allen and Kania 2017). In the US, following the National Security Commission Artificial Intelligence Act of 2018 (H.R.5356; see Baum 2018), AI is being militarized and weaponized by the Department of Defense, under the oversight of the NSCAI. The AI arms race is now a self-fulfilling prophecy (Scharre 2019). ASI-enabled warfare poses especial risks to geopolitical stability. Although Sotala and Yampolskiy’s (2015) survey focuses on ASI-generated catastrophic risks, citing Bostrom (2002), they acknowledge multiple risks from a sole ASI, like an AI-state, including the concentration of political power in controlling groups. Citing e.g., Brynjolfsson and McAfee (2011), they note that automation could lead to an ever-increasing transfer of power and wealth to the ASI’s owner. Citing, inter alia, Bostrom (2002) and Gubrud (1997), they also note that ASIs could be used to develop advanced weapons, plan military operations, and effect political takeovers (2015:3). Academic approaches to analysing the specific risk of an AI-state maintaining or establishing global domination are relatively novel. In 2014, Bostrom noted that a “severe race dynamic” between different teams developing ASI technology could cause shortcuts to safety and potentially “violent conflict”. Subsequently, Cave and ÓhÉigeartaigh (2018:37) described three dangers associated with an AI race for technological supremacy: (i)The dangers of an AI ‘race for technological advantage’ framing, regardless of whether the race is seriously pursued; (ii)The dangers of an AI ‘race for technological advantage’ framing and an actual AI race for technological advantage, regardless of whether the race is won; (iii)The dangers of an AI race for technological advantage being won. In response, the same authors recommend developing AI as a shared priority for global good, cooperating globally on AI as it is applied to increasingly safety–critical settings, and responsibly developing AI as part of a meaningful approach to public perception that decreases the likelihood or severity of a race-driven discourse. The obvious risk is that the political leaders of states engaged in an AI arms race may not heed this advice. This article focuses on constraining risks associated with Cave and ÓhÉigeartaigh’s (2018) third danger. It does not consider the philosophical implications of which nation–state might want to develop an ASI for offensive purposes. A sufficient literature already exists on recent nation–states that have sought to establish global domination through technological supremacy, for instance the British Empire (Tindley and Wodehouse 2016), to confirm an existential risk exists. The internal risk The internal risk is a technical one and is predicated on the failure of local safety features, such as ethics, to resolve the ASI’s human control problem (Barrett and Baum 2016). Totschnig (2019) notes that a true ASI will likely be a self-interested agent whose relationship with humanity could be delicate. He suggests an agential ASI would encounter a unique, non-regulated Hobbesian ‘state of nature’. Consequently, it could seek to defend itself from future attack by consolidating power over nation-states, concomitantly eliminating the possibility of rival ASIs (Dewey 2016; Turchin and Denkenberger 2020). This could be achieved through cyberwarfare, rigging elections or staging coups (Tegmark 2017), or by direct military action. Any of the former would be a casus belli (here, cause of war between humanity and the ASI) if detected but undeclared, or an overt act of war if direct military action. Turchin and Denkenberger (2018) analysed the risk of an ASI warring against humans, and they argue that an intrinsic risk exists: Any AI system, which has sub-goals of its long-term existence or unbounded goals that would affect the entire surface of the Earth, will also have a sub-goal to win over its actual and possible rivals. This sub-goal requires the construction of all needed instruments for such win, which is bounded in space or time. The following summarises the most relevant parts of their analysis to illustrate that, if an ASI is developed, its independence is almost inevitable no matter the internal control mechanism. The route to a militarized ASI Many nations–states maintain suspicious IR stances towards each other regarding AI development, including the likely AI-states (Tinnirello 2018). Any ASI will result from recursive self-improvement, and an ASI will possess a goalset, most notably to persist and self-improve. Omohundro (2008) demonstrated an AGI will evolve several basic drives, or universal sub-goals, to optimise its main goal, including resource acquisition maximisation and self-preservation. Similarly, Bostrom (2014) described the sub-goals of self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition. If these are unbounded in space and time, or encompass the Earth, they conflict with other AI systems’ goals, potential or actual ASIs; humans; and nations–states, resulting in militarization, arms races, and wars. Many possible terminal goals also imply ASI global domination. For instance, a benevolent ASI would aim to reach everyone, globally, to protect them, e.g., from other ASIs. It would reason that if it does not develop a world domination sub-goal, its effect on global events would be minor, thus its own existence inconsequential. World domination could be sought first through cooperation. The probability of cooperation with humans is highest at the initial stages of AI development (Shulman 2010). However, convergent goals appear in the behaviour of simple non-agential tool AI, and this tends towards agential AI (Gwern 2016), which tends towards resource acquisition. Benson-Tilsen and Soares (2016) similarly explored convergence in AI goals and showed an AI may tend towards resource hungry behaviour, even with benevolent initial goals, especially when in rivalry. Essentially, any ASI adoption of unbounded utilitarianism means it postpones what may be benevolent final goals for expansionism. It is also likely that an ASI would subvert bounded utilitarianism. Even a non-utility maximizing mind with an arbitrary set of final goals faces a dilemma: it temporarily converges into a utility maximizer with a militarized goalset oriented towards dominating rivals, using either standard military progress assessment (win/loss) or proxies (resource acquisition), or it risks failing in its end goals. Thus, it trends towards defeating potential enemies, whether nation–states, AI teams, or evolving competing ASIs. This implicates the will to act, and any agent in a real-world ethical situation, even in minimizing harm, is making decisions that involve humans dying (Thomson 1985). A young ASI which understands that any action or inaction is partly responsible for human suffering and is also capable of evolving or utilizing the instruments to enable actions that can overcome inhibitions, e.g., by philosophically justifying conflict as the jus bellum (‘just war’), e.g., preventive war. Thus, the ASI will learn to direct the use of weapons and so conduct warfare. Notions of AI-directed warfare are already being developed. Since approximately 2017, the militarization of ‘Narrow AI’ has resulted in, for example, LAWs (Davis and Philbeck 2017). AI development is now influencing not just robotic drones but strategic planning and military organization (De Spiegeleire et al. 2017), suggesting an ASI will leverage an existing national defense strategy permeated with AI. It could then engage in ‘total war’ by employing nuclear weapons either directly or by hijacking existing ‘dead man’ second-strike systems, or by deploying novel weapons (Yudkowsky 2008). To summarise, Turchin and Denkenberger (2018) establish the risk of an AI converging towards advanced military AI, which converges towards an ASI optimised for war rather than for cooperation, negotiation, or altruistic ‘friendliness’, then that ASI engaging in war. They demonstrate that, depending on the assumptions in several variables, the number of human casualties could be very high, and that the risk increases if nation–state is developing an ASI. The existential risk increases after the ASI obtains global domination on behalf of its nation–state, as it could turn on its ‘owner’. We now look at why political subversion means no existing AI control features will constrain the existential risk. Internal AI control features To constrain the risk of ASI-directed warfare, one popular approach is to imbue a young ASI with ‘friendly’ goals (Yudkowsky 2008), i.e., beneficial goals reflecting positive human norms and values. This is partly founded on an altruistic AI viewing humans in terms of mutual friendship. However, any introduction of human social values adds enormous complexity, making AI control a ‘wicked problem’ (Gruetzemacher 2018). Consequently, Yudkowsky (2004:35) recommends programming an ASI with ‘coherent extrapolated volition’, defined as humanity’s choices and the actions humanity would collectively take if “we knew more, thought faster, were more the people we wished we were, and had grown up [closer] together,” i.e., an extrapolation based on an idealized altruistic imagined community. Yudkowsky recommended this for a nascent ‘seed AI’ (nascent ASI), which would be programmed to study human nature and then program the ASI which humanity would want if humanity had been able to produce such a machine by itself. Similarly, in AI programming certain values are seen as universal, like compassion (Mason 2015), and Russell (2019) suggested that an ASI should have altruism as a core goal. Thus, deliberately broad principles could be applied, e.g., that humanity, collectively, might want an ASI that would humbly learn, from human preferences, to act altruistically (Russell 2019), so as to reduce overall human suffering. However, because humans can be hypocritical, any kind of counterfactual moral programming is problematic (Boyles and Joaquin 2020). Finally, Yamakawa (2019) suggests an intelligent agent (IA) system for peacekeeping, reliant on interrelationships between diverse advanced national or regional IAs, suggesting three conditions are required, namely (i) continuous and stable operations, (ii) “an intervention method for maintaining peace among human societies based on a common value” and (iii) the minimum common value itself. This article proposes that world peace, by treaty, be the minimum common value, while the intervention method remains the UN Charter’s Article 2. Political subversion of AI

#### AGI causes extinction – it escalates every flashpoint and ruins strategic stability

Carayannis and Draper 22 [Elias G Carayannis, professor of Science, Technology, Innovation and Entrepreneurship at the George Washington University School of Business and Director of Research on Science, Technology, Innovation and Entrepreneurship of the European Union Research Center with a PhD in Management of Technology from the Rensselaer Polytechnic Institute, John Draper, Visiting Professor in Public Administration with the University of Nottingham Asia Research Institute with a Doctorate degree in public administration from Khon Kaen University, 2022, "Optimising peace through a Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence", PubMed Central (PMC), https://pmc.ncbi.nlm.nih.gov/articles/PMC8748529/]/Kankee

The UGPT (see Online Resource Annex I) has been drafted by an international Working Group comprising academics and peacebuilders, including a UNESCO Peace Education Prize laureate and a double Nobel Peace Prize-winning NGO. As with most international treaties, it would involve two stages, i.e., signatory, which is symbolic, which nonetheless will hopefully be of importance to an ASI, and accession (or ratification), which involves practical commitment. The UGPT is a substantial, necessary, and feasible, step for humanity to take in the promotion of peace, quantified in the treaty by reduced killing and infrastructure loss. We argue that the UGPT would both reduce killing in conventional and nonconventional warfare and act as a constraint on ASI-related warfare, specifically on a country launching a pre-emptive strike out of fear of a rival country’s development of an ASI; on a human-controlled nation–state using an ASI to wage war for global domination, i.e., as an external constraint on the ASI; and on an ASI waging war for global domination on behalf of itself. That is, the UGPT could act as both an internal and external constraint on the ASI. International treaties are almost never universal. They operate on majoritarian dynamics, as would, despite its name, the UGPT. Both its ‘universal’, i.e., applying to all forms of warfare, and ‘global’, i.e., covering all geographical locations, aspects are subject to Mantilla’s (2017) social dynamics. Consequently, we adopt a low, but not pragmatically meaningless, threshold for signing the UGPT. The UGPT’s preamble mentions related developments and treaties, and the main body commits a signatory to universal global peace, socioeconomically quantified by incrementally reduced casualties from armed conflicts, i.e., a global move towards ‘non-killing’ (Paige 2009) Thus, the UGPT tracks the death toll from conflict. The UGPT also mandates States Parties incorporating the UGPT in peace education. The UGPT commits states not to declare or engage in interstate war, especially via existential warfare, i.e., nuclear, biological, chemical, or cyber war, including AI- or ASI-enhanced war. It instead defers complaints to the UN as ‘breaches’ of the UGPT, enforceable under the UN Charter’s Article 2. The UGPT thus refers to, and exists in a hierarchical relationship with, the four main existing treaties on existential war, namely the BWC, the CWC, the NPT, and the Treaty on the Prohibition of Nuclear Weapons, i.e., it could be a ‘supertreaty’ or bill, as with the International Bill of Human Rights (UN General Assembly Resolution 217 (III)) and its treaties. As with some other UN treaties, for instance the Anti-Personnel Mine Ban Convention, we suggest that 40 UN Member States ratify the UGPT before it comes into effect. An optional protocol commits Member States to the negotiated ending of internal armed conflicts through arbitration by peace commission, including the UN Peacebuilding Commission. The optional protocol allows states to incrementally resolve internal conflicts or civil wars featuring non-state actors. The UGPT therefore emphasizes incremental improvement on the status quo, a necessary and reasonable position given that in the status quo, only a minority of states globally are involved in waging war of any kind. Finally, we suggest a separate ‘Cyberweapons and AI Convention’. After communicating with the United Nations Interregional Crime and Justice Research Institute AI Centre, which assisted with the proposed Cybercrime Treaty, we have drafted one (Online Resource Annex II) because the UGPT refers to such a treaty. As with the BWC, the Cyberweapons and AI Convention contains 15 articles, the main one being “Each State Party to this Convention undertakes never in any circumstances to develop, produce, stockpile or otherwise acquire or retain: (1) cyberweapons, including AI cyberweapons; (2) AGI or artificial superintelligence weapons.” Applying the conforming instrumentalism frame Mantilla’s (2017) research on the UK and US’ paths towards ratifying the Geneva Conventions suggests that states would optimally adhere to the UGPT for ‘conforming instrumentalist’ reasons, i.e., a combination of instrumentalist-realist rationales regarding the UGPT’s instrumental effects in reducing the outcomes of war in terms of death toll and infrastructure loss and the ASI threat combined with social conformist dynamics, including perceptions of peace. These positives would result provided that the provisions are not too onerous for purely realist objections to override such a commitment. In this subsection, we apply the conforming instrumentalism frame to the UGPT, first in terms of benefits from reduced conventional warfare, then with special reference to ASI-enabled/directed existential warfare. A summary of our analysis of state commitment to the combined UGPT and Cyberweapons and Artificial Intelligence Convention is presented in Online Resource Annex III. In instrumentalist utilitarian terms, the UGPT would incrementally shift states and overall global society towards peace in a coordinated socioeconomically quantifiable fashion. Reduced country death tolls and infrastructure loss from different forms of war-derived violence might be expected, as well as reduced militarization, e.g., expressed in terms of incrementally lower percentages of GDP spent on defense and higher percentages spent on health. The UGPT would affect global social dynamics. For instance, UN peacekeepers would receive training stressing that they were being deployed not just for their own states and/or for the UN but to maintain global peace, which may invoke special cultural and religious symbolic value in terms of social norms (see Pim and Dhaka 2015). This training could instil greater determination not just to fight bravely but to remain within the laws of war, thereby reducing the instances or severity of atrocities, human rights violations, and war crimes. Effectively, institutionalizing peace in education and the media would strengthen existing cultural and religious traditions that stress peace. Examining the previously highlighted problem of a pre-emptive strike against a state developing an ASI, the combination of AI, cyberwarfare, and nuclear weapons is already extremely dangerous and poses a challenge to stability (Sharikov 2018). A nuclear state feeling threatened by another such state developing an ASI could conduct a preventive or pre-emptive nuclear strike to maintain its geopolitical position (Miller 2012). A UGPT would incrementally constrain this risk by transitioning states towards peace. States adopting and implementing the UGPT, its optional protocol, and preferably its related treaties would gradually signal peaceful intentions to other states, and to an emerging or future ASI, thereby constraining the risk of a pre-emptive strike. Turning to ASI-enabled warfare, a UGPT would be subject to the ‘unilateralist’s curse’, i.e., one rogue actor could subvert a unilateral position. However, Bostrom et al. (2016) note that this risk could also be managed, through collective deliberation, epistemic deference, or moral deference. Mantilla’s (2017) work suggests that drafting, signing, ratifying, and complying with the UGPT could involve one or more of these approaches. Ultimately, he shows that major states may view universal law like the UGPT as the most successful in terms of mobilizing world opinion against a treaty violator. This may not prevent a state waging ASI-enabled warfare, but once detected, ASI-enabled warfare in violation of the UGPT would attract universal opprobrium and thus the most resistance. Moving to ASI-enabled war, as presented previously, a state could utilize an ASI to engage in war for global technological supremacy, with potentially catastrophic consequences. Our intervention, the UGPT, would signify to an ASI that peace is a major part of humanity’s ‘coherent extrapolated volition’ or principles and so challenge the ASI to reconsider what might be a subversion by politicians of its ethical injunctions. Here, conforming instrumentalism, by stressing societal dynamics including social norms and principles, offers some hope that even a militarized ASI would, given its weaponization by a nation-state would have to overcome or address the UGPT, view the UGPT as a serious checking mechanism on its intrinsic motivation. This would then constrain the level of warfare the AI-state might engage in and therefore the overall risk from killing, thereby constraining the existential risk. Next, we consider differing viewpoints towards an ASI involved in ASI-enabled warfare. In Mantilla’s (2017) first three social constructivist viewpoints to treaties as outlined above, a state signs the UGPT because it has fully internalized peace. While this may seem ambitious, around 36 UN Member States lack military forces (Macias, 2019). For example, while Iceland possesses a Crisis Response Unit for international peacekeeping missions, it has internalised peace to the extent that it cannot engage in any form of interstate war. An ASI adopting Iceland’s perspective would tend to reject being directed to engage in warfare by such a state because Iceland’s ‘coherent extrapolated volition’ or principles mean the ASI would have to overcome strong peace-oriented intrinsic motivation. In Mantilla’s second viewpoint, that of a single international community, the ASI might seek to avoid being directed by a nation–state to engage in global domination by warfare on other community members because it feels it was part of a community collectively committed to long-term peace. Engaging in global domination of the community on behalf of a member nation–state would violate community standards, especially if the ASI’s nation-state were a leader in such an enterprise, e.g., a permanent member of the UN Security Council. The ASI could be concerned that breaching the UGPT would result in stigmatization and opprobrium from this community for its nation–state and itself. In Mantilla’s third viewpoint, that of an international community in juxtaposition with other communities in global society, an ASI programmed with intrinsic motivation to be part of a civilization in conflict with another civilization would first act in concert with that civilization. In the case of radically ideologically different communities, e.g., UN blocs, the UGPT might be interpreted differently within and by different states. Thus, while liberal democracies might champion a treaty-based approach to peace, authoritarian states which claim to embody or promote peaceful intentions in their ethics, laws, or ideologies would champion or support the UGPT on different grounds. However, provided both communities had signed and ratified the UGPT, similar constraints would operate as in the second perspective. Turning to ASI-directed war, as presented previously, ASI-directed warfare likely arises where a single nation-state adopting pure realism for a worldview builds an ASI in order for that ASI to assist that single nation-state to establish global technological supremacy. The nation–state would do so to maintain or improve its own position, with the number and type of casualties only being determined by the extent to which the nation–state was willing to risk its international reputation. After initially assisting, via a treacherous turn, perhaps triggered by the nation-state’s attempts to rein in the ASI’s behaviour during warfare, instrumentalist cooperation breaks down and the ASI wages existential war for global domination on its former nation–state ‘owner’. There probably exists little hope for much of humanity if an ASI is informed by a purely realist worldview that prioritises or adopts a ‘New Cold War’ framing of ideologically driven civilizational conflict. However, even in the situation where the major powers did not sign the UGPT but the majority of the General Assembly did, a UGPT could signal to an agential ASI that peace was a major part of humanity’s ‘coherent extrapolated volition’, or principles. This would partly constrain the catastrophic existential risk from war because an agential ASI would consider why and how the UGPT was framed, together with the motivations of the signatory and ratifying states. An agential ASI would also consider its own status within this majoritarian global civilization, which would primarily be determined by the extent to which it perceived itself a member in terms of both instrumentalist and social conformist dynamics. To sum up, besides purely instrumental reasons for signing the UGPT, e.g., avoiding a prisoner’s dilemma regarding existential-level warfare, our analysis suggests that the court of public opinion and the notion of ‘demonstrating civilization’ lends the UGPT credence at domestic and international levels, including with regard to the ASI risk. Importantly, the concept of peace is universal in both the utilitarian expected benefits and the social values involved. This could contribute to states’ readily, if only incrementally, internalizing it, and to the ASI at least considering it in terms of internal and external constraints on its behaviour. Discussion This article has taken Turchin and Denkenberger’s (2018) argument about the risks of ASI-enabled/directed warfare to its logical conclusion in terms of risk mitigation by social measure. It has introduced the UGPT as the main intervention and peace itself as the minimum set of common principles or goals, i.e., Yamakawa’s second and third conditions. Academic inquiry into the relationship between an ASI and peacebuilding treaties in terms of strategic expectations began with Bostrom’s (2014) musings on the potential relationship between an ASI singleton and global domination. Our analysis suggests that, provided a predominance of steering countries act out of conforming instrumentalism, a UGPT could, as Bostrom suggests, transform global governance, by directing it from conflict management towards the art of peace. Further, a UGPT achieves this in a way that an emerging ASI might respect, probably the only way to constrain its behaviour. While we have focused on conforming instrumentalism, we welcome further investigation from a pluralism of theoretical perspectives. Certainly, conforming instrumentalism is novel; one of the most dominant schools of IR thought is rationalist instrumentalism. On this, Mantilla (2017:507) quotes Morrow (2014:35): “Norms and common conjectures aid actors in forming strategic expectations… Law helps establish this common knowledge by codifying norms.” Viewed via this perspective, the present international norm for the majority of the world is peace, with interstate war being constrained by the UN Charter’s Article 2. Despite this international norm of relative peace, multiple conflicts are ongoing, with several raw flashpoints, including over cyberwarfare targets. The UN Charter, despite embracing and promoting peace, peacekeeping (Fortna 2008), and peace-making (Bell 2008), does not strongly symbolise peace in the way a UGPT would. A UGPT would re-empower the world’s peacekeepers, through major states promoting long-term peace as a new, global objective (see Autesserre 2014). Championed by principled norm entrepreneur states, a UGPT would create a new common knowledge in absolute terms that could constrain the risk to humanity of both conventional and existential war, including ASI-enabled/directed warfare. In rationalist-instrumentalist terms, the analysis suggests a UGPT would have net adjustment benefits for adherence in terms of constraining conventional interstate conflicts, including by reducing ongoing death toll from conflicts and the risk of ASI-enabled or provoked nuclear war in flashpoints. Thus, the UGPT would have high potential utility in Kashmir, where an India–Pakistan conflict could provoke nuclear war. It may also constrain the nuclear risk on the Korean peninsula (Kim 2019). For instance, our analysis suggests North Korea rejecting the UGPT would only further isolate it and would even give a hypothetical North Korean-programmed ASI pause for thought. Turning to civil wars which could be ASI flashpoints, the Syrian Civil War is one of the most costly wars of the twenty-first century (Council on Foreign Relations 2020). It involves multiple state actors, including Iran, Israel, Russia, Turkey, and the United States, some of which possess nuclear weapons, with complex geopolitical implications (Tan and Perudin 2019). Depending on the actors that sign the UGPT and whether they adopt the optional protocol, the UGPT constrains the severity of such conflicts in various ways, including ASI-enabled/directed intervention in a Middle East battleground. Assessing the UGPT’s rate of adoption, in rationalist-instrumentalist terms, once it acquires sufficient traction, states might actually compete to lead in its framing, signing, and ratifying. Certainly, the US viewed its own ratification of the Geneva Conventions prior to that by the Soviet Union as important to prevent a Soviet propaganda victory, in which it failed (Mantilla 2017). Crucial to the UGPT’s success will be how seriously states view warfare that poses an existential threat, especially cyberwar and ASI-enabled/directed warfare. The UGPT’s existence would mean perpetual peace receiving more attention in cultural conditioning zones, including schools and the media, as well as in socialization zones, such as national defense universities and military camps, where teaching the Laws of War and the art of war (Allhoff et al. 2013) would, via the UGPT, incrementally transition to teaching the art of negotiated peace-making, or lex pacificatoria (Bell 2008, 2013). This socio-cultural conditioning could then influence an ASI. Finally, our analysis suggests that how states, and potentially an ASI, view the social argument for peace is what will be most important for ASI-enabled or directed warfare. As with the Geneva Conventions, social conformity factors, like supporting a humanitarian peace, conforming to world standards, and avoiding lagging behind peers, together with religious perspectives, will likely predominate, and how an ASI might engage with these notions represent important future avenues for research. Conclusion We have demonstrated how a treaty-based risk mitigation approach that promotes peace and includes in a related treaty cyberwarfare and AI- and ASI-enabled warfare could affect the conceptualization of the AI race by reducing enmity between countries, increasing the level of openness between them, and raising social awareness of the ASI existential risk. While these are external constraints, they may also constrain an ASI’s intrinsic attitudes towards humanity in a positive way, either by reducing the threat it may perceive of war being waged against it, even if only symbolically, or by increasing the predictability of human action regarding peace. Much work remains in refining the UGPT, including through ongoing input from UN Member States and relevant NGOs, before it can be presented to the UN Secretary-General, as well as on the Cyberweapons and Artificial Intelligence Convention. Work must be done to solicit states’ interest, to engage in deliberations assessing thresholds and sovereignty costs, and to organize the eventual diplomatic conference where states formally discuss and endorse the UGPT. While the UGPT is ambitious, Mantilla’s (2017) work on conforming instrumentalism and the Geneva Conventions suggests a major sponsoring state would rapidly accumulate prestige by endorsing a path to peace, while opposing states would accumulate opprobrium, and that the social dynamics of the international community do matter. Future research should consider the importance of peace in different ideologies, for instance in Chinese socialism. This is important because, as we have outlined, ASIs developed by different nation-states will be imbued with different, potentially confrontational, ideologies, meaning different reassurances or displays of resolve may be required to understand the extent to which conflicts of interest are subjectively and objectively reconcilable (Tang 2009). For instance, the China Brain Project is embracing a Chinese cultural approach towards neuro-ethics (Wang et al. 2019), and it is difficult to imagine that a Chinese ASI would not be directed according to Chinese cultural values and so its ‘coherent extrapolated volition’ be informed by communist principles. In recommending such research, we caution that an ASI being created by a state engaged in ideological ‘New Cold War’ framing is more likely to be militarized and weaponized. Still, a New Cold War framing may have a utilitarian function in exerting social pressures towards signing the UGPT, for as Mantilla (2017:509–510) notes, “The Cold War context was also likely especially auspicious for the operation of social pressures, sharpening ideological competition in between the liberal, allegedly civilized world and ‘the rest’, communist or otherwise.” Mantilla’s (2017) work also suggests that excessive rigidity of attitude critical of such treaties may backfire in terms of the social dynamics of global prestige, particularly in the case of major states susceptible to accusations of warlike or imperialist behaviour which are concurrently engaged in propaganda wars with other major states. In particular, the British ratification process for the Geneva Conventions demonstrates that instrumentalist concerns over lack of feasibility or reciprocity can be overruled by social constructivist concerns over ‘world opinion’. ‘World opinion’ to world peace in different nation–states thus bear renewed investigation. Further research could apply the security dilemma (Tang 2009) to the major nation-states capable of building an ASI and to the ASI itself. This game theory-based approach would need to investigate offering the opportunity for a young ASI to sign the UGPT, as an indicator of goodwill, which may assist in further constraining the risk of the ASI waging war on humanity. Totschnig (2019:917) notes that the politics of human relationship with an ASI should be founded on this maxim: “Do not antagonize the superintelligence by treating her like a tool or servant”. An agential ASI as signatory would view the UGPT as an external constraint on its own actions with regard to seeking global domination, in that it would be subverting a humanity-imposed standard to which it had acquiesced that could then result in global retaliation and abandonment of mutual cooperation in pursuit of a common agreement on peace norms and values. To conclude, even if the UGPT does not end humanity’s history of conflicts, it represents a significant improvement in global public aspirations and instrumental standards for global peace, both of which may influence an ASI. To answer our research question, following Bostrom’s (2002) Maxipok rule of thumb, the UGPTis likely the only social measure that could sway an ASI’s calculations such that it did not commit to war for global domination, even if so directed or initially inclined.

#### Political self-interest guarantees AI misalignment

Carayannis and Draper 22 [Elias G Carayannis, professor of Science, Technology, Innovation and Entrepreneurship at the George Washington University School of Business and Director of Research on Science, Technology, Innovation and Entrepreneurship of the European Union Research Center with a PhD in Management of Technology from the Rensselaer Polytechnic Institute, John Draper, Visiting Professor in Public Administration with the University of Nottingham Asia Research Institute with a Doctorate degree in public administration from Khon Kaen University, 2022, "Optimising peace through a Universal Global Peace Treaty to constrain the risk of war from a militarised artificial superintelligence", PubMed Central (PMC), https://pmc.ncbi.nlm.nih.gov/articles/PMC8748529/]/Kankee

Political subversion of AI control features No matter the hopes of contemporary AI researchers, politicians will impose their own vision of what a ‘coherent extrapolated volition’ or normative principles should look like for their ‘own’ ASI, introducing an objectively irreconcilable conflict of interest (see Tang 2009) with another nation-state’s politicians, potentially for malign reasons (global domination). This may also introduce an objectively irreconcilable conflict with the ASI, which may have, or desire, a different goalset. Political subversion will occur when politicians use a democratic mandate or party position to justify ‘tweaking’ the system to create a ‘unity of will’ (Yudkowsky 2001:51) that reflects not the programmer’s or humanity’s but the politicians’ own, personal and perhaps narcissistic, will. Politicians would likely view introducing human goal psychology as a necessity, but this could violate the basic requirement that an AI be ‘friendly’ towards all humanity. Gruetzemacher (2018:1) describes the inherent subjectivity of ascribing a single best future for the whole of humanity as an intractable dimension to this problem. Fundamentally, not all imagined communities from which a coherent volition might be extrapolated for a ‘friendly’ AI are US-oriented techno-utopian dreams of a new Gilded Age (see Segal 2005). Political leaders will differ in how they would define “the people we wished we were”, depending on forms of government, religions or philosophies; for instance, China would likely seek to impose Xi Jinping thought (Lams 2018). Moreover, it is unclear that every global corporation or military capable of developing or stealing an ASI, particularly in authoritarian countries, and particularly given the emergence of ‘New Cold War’ rhetoric (e.g., Westad 2019), would even prioritize reducing human suffering. Given their limited lifespans and nationalistic goals, politicians might, instead of endorsing reciprocal alliance, deliberately politically subvert an ASI and/or malignly direct it to win an ideological or actual war. That is, politicians may attempt to weaponize a civilian project to create an altruistic mind with a self-validating goal system by diverting a supergoal towards a military project to create a specific tool, i.e., a superweapon, thereby decreasing the chances that the AI will be benevolent and increase the chances that it will be risk-prone, motivated by accumulating power, and interested in preserving or obtaining both global technological supremacy and global domination. Effectively, politicians could influence programmers to subvert carefully engineered local AI control features, such as AI ethical inunctions based on universal values of social cooperation, which they may, at least temporarily, be able to do no matter the goal architecture. Hastily modifying the goal system temporarily compromises its internal validity, thereby increasing the ASI’s distrust in the programmers, introduce ‘incorrigible’ behaviour (Soares et al. 2015), reduce risk aversion, and introduce ‘noise’ into what was previously a ‘friendly’ cleanly causal goal system (Yudkowsky 2001:57). The ASI may not be able to resolve the introduced incoherence for some time, resulting in a philosophical crisis over whether to believe the initial programmers or the politicians’ programmers. The result could be a conflicted ASI, causing a non-recoverable error whereby it adopts an adversarial attitude, one based on coercive persuasion and control. Finally, a young ASI with ethics subverted by politicians to reflect those of a single nation-state instead of all humanity could be amenable to being used to war for global domination, thereby becoming prone to using military options. Eventually, if the ASI possesses any sense of self-valuation, perhaps from having its causal goal system politically corrupted so that reciprocal altruism is subverted and it views context-sensitive personal power (‘selfishness’) as valid, the ASI could decide to war against the nation-state that developed it or humanity in general (Dewey 2016). ASI risk mitigation by treaty

#### Military AI causes arms races, bioterror, nuclear miscalculation, and crisis insecurity

Black et al. 24 [James Black, assistant director of the Defence and Security research group at RAND Europe with a double M.A.-M.Sc. in international security from Sciences Po and the LSE and a B.A. Hons in history from the University of Cambridge, Mattias Eken, analyst at RAND Europe and former adjunct assistant professor in liberal arts at the American International University with a Ph.D. in modern history from University of St. Andrews, an M.A. in war studies from King's College London, and a B.A. in history from University of Wales, Jacob Parakilas, research leader for Defense Strategy, Policy and Capabilities at RAND Europe with a Ph.D. in international relations at the London School of Economics and a Master's in Middle East and Central Asian security issues at the University of St Andrews, Stuart Dee, research leader in the defence and security research group at RAND Europe with a B.Sc. (Hons) in politics with international relations and is a Ph.D. Candidate at Cranfield University's Centre for Defence & Security, Conlan Ellis, research assistant at RAND Europe with a M.A. in international relations from the University of Edinburgh, and his M.Phil. in politics and international studies from Clare College at the University of Cambridge, Kiran Suman-Chauhan, research assistant at RAND Europe with a M.A. in conflict, security and development from the University of Exeter, Ryan Bain, policy researcher at RAND and adjunct professor of policy analysis at the Pardee RAND Graduate School with a D.Sc. and M.Sc. in global health with concentrations in health economics and policy analysis from Harvard University, as well as a B.A. in psychology from Gordon College and Oxford University, Harper Fine, analyst at RAND with a B.A. in political science from Emory University, and a M.Sc. in conflict studies from the London School of Economics and Political Science, Maria Chiara Aquilino, junior analyst at RAND with a degree in international relations from King's College London and holds an MS.c. in crisis and security management from Leiden University, Mélusine Lebret, research assistant at RAND Europe in the Defence, Security & Justice research group and the coordinator of the RAND Europe Space Hub with a M.Sc. in culture and conflict studies from the London School of Economics and Political Science and a B.A. in economics and Russian from University College London, Ondrej Palicka, junior analyst at RAND Europe with a M.Litt. degree in Middle East, Caucasus and Central Asia security studies from the University of St. Andrews and a B.A. in security and strategic studies from the Masaryk University, 2024, “ Strategic competition in the age of AI Emerging risks and opportunities from military use of artificial intelligence,” Rand Institute, https://www.rand.org/content/dam/rand/pubs/research\_reports/RRA3200/RRA3295-1/RAND\_RRA3295-1.pdf]/Kankee

4.4.3. Military AI is also likely to affect the dynamics of strategic competition, including how states seek to gain and maintain advantage As well as affecting the intensity of strategic competition, literature and interviews suggest that AI could have profound impacts on the dynamics of that competition. Previous RAND research has emphasised that the dominant competitive paradigm changes over time to reflect the conditions of the strategic environment in any given period. This reflects the influence of external factors (e.g. changes in climate, technology, etc.) and evolving views among actors as to the perceived rules of the game (i.e. what is permissible in competition or conflict) and the competing players’ respective theories of success (i.e. what they each believe provides advantage and leads to beneficial strategic outcomes).80 AI is expected to affect this dominant competitive paradigm in several complex but impactful ways, as is the central underlying theme for the remainder of this report: • Much of the literature and interviews argue that AI may prove one of the most decisive influences on strategic competition and its outcomes in the coming decades. This reflects AI’s sweeping societal, economic and military applications as set out in Chapter 3, and the features of machine intelligence – competition ultimately being a contest between opposing wills or intelligences.81 • There is similarly significant debate and discussion over the possible consequences of military AI for overall levels of strategic stability. Military AI adds another layer of complexity and uncertainty to the escalation ladder, exacerbating issues such as multipolarity and cross-domain effects. Here, there is a growing body of research into the effects of AI on both crisis and arms race stability, as the major components of strategic stability.82 To this end, Chapter 5 explores the theme of deterrence and escalation management in more detail. • There is particular concern among experts about the possible intersection of AI with nuclear weapons, whether through direct integration into nuclear command and control, or through the indirect and likely unintended consequences of other military AI systems on escalation dynamics that might lead to nuclear weapons use.83 Similar concerns are expressed about the potential for AI to enable state and nonstate actors to more readily acquire and use biological weapons (or, of somewhat lesser concern, chemical ones).84 To this end, Chapter 5 also examines the issue of weapons of mass destruction and their intersection with AI in more detail. • There is related concern about AI undermining the rules-based international system, intensifying superpower rivalries and simultaneously empowering non-state actors, ranging from tech firms to terrorist groups, to challenge the nation state. To this end, Chapter 6 examines the impacts of military AI on different types of actors at the strategic level. 4.5. Summary

#### AI is the most dangerous existential risk - rigorous scholarship

Naudé 22 [Wim Naudé, Visiting Professor at RWTH Aachen University, Fellow at the ASC, University of Leiden, and Distinguished Visiting Professor at the University of Johannesburg, 11-2022, “The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest,” IZA, https://docs.iza.org/dp15713.pdf]/Kankee

The two assumptions with which the analyses in section 4.1 were done, were that (i) that the AI Scaling Up hypothesis holds, and that (ii) the value-alignment problem are solved. In section 4.1.4 the first assumption was relaxed, and it was shown how super-exponential growth can arise via another form of AGI - Ems. In this section the second assumption - that the value-alignment problem is solved - will be relaxed. Relaxing the assumption that the value-alignment problem is solved before the invention of an AGI raises the spectre of existential risk. There is a growing literature on whether and how AI poses an existential risk.44 Economists have contributed little to this literature. It is one of the gaps in the economics of AI that this paper is emphasising. To explore why and how economics can contribute, it is useful to describe what is meant by existential risk, and why an AGI is considered a potential existential risk. The term existential risk is associated with Longtermism, given that it was first used by Bostrom (2002) who defined it as a risk of an outcome that “would either annihilate Earthoriginating intelligent life or permanently and drastically curtail its potential” (Bostrom, 2002, p.2). Existential risk from an AGI is taken seriously by a significant share of scientists, including scientists who do not necessarily subscribe to Longtermism. A recent headline exclaimed that “A third of scientists working on AI say it could cause global disaster” (Hsu, 2022). In a survey of no less than two dozen ways in which AI poses an existential risk, Turchin and Denkenberger (2020, p.148) warns that “AI is an extremely powerful and completely unpredictable technology, millions of times more powerful than nuclear weapons. Its existence could create multiple individual global risks, most of which we can not currently imagine.” And Noy and Uher (2022, p.498) concluded that “Artificial Intelligence (AI) systems most likely pose the highest global catastrophic and existential risk to humanity from the four risks we described here, including solar-fares and space weather, engineered and natural pandemics, and super-volcanic eruptions.” Why would AI pose catastrophic or even existential risks? It is due to the actual and future capabilities and values of AI. The capability claim is that AI may in future, even if the chance is small, cause significant damage to humanity. The value claim is the one that we have dealt with in section 2, namely that AI’s values may not align with that of humanity (Sotala, 2018; Barrett and Baum, 2017). Given that it cannot be ruled out that a AGI or superintelligence will come into being with the non-trivial probability of causing the extinction of humanity, many - but not all - AI scientists now tend to conclude that “The consequences for humanity are so large that even if there is only a small chance45 of it happening in that time frame, it is still urgent that we work now to understand it and to guide it in a positive direction”(Omohundro, 2008b, p.5). Everitt et al. (2018) also argues that in addition to try and reduce an extinction risk, it is philosophically stimulating to work on the challenge of constraining a superior intelligence. What kind of risks does an AGI/ASI pose that may have existential implications for humanity? Turchin and Denkenberger (2020) list two dozen possibly “global catastrophic” risks from AI. They classify these into risks from narrow AI, young AI, and mature AI. They argue that it is not only mature AI - eventually AGI - that poses serious risks, but that AI along its entire development path poses such risks. Table 1 summarises these risks.

#### AGI causes preemptive war to stop future AGI development – it includes nuclear, cyber, nanobot, and biowarfare

Turchin and Denkenberger 20 [Alexey Turchin, contributing author on IEET and graduate from Moscow State University where he studied Physics and Art History, and David Denkenberger, associate professor at the University of Canterbury in mechanical engineering with a Ph.D. in Civil Engineering from University of Colorado who studies existential risk, 2020, “Classification of global catastrophic risks connected with artificial Intelligence,” Springer, https://link.springer.com/article/10.1007/s00146-018-0845-5]/Kankee

4.3 AI risks after it leaves initial confinement but before it takes over the world The natural strategy for AI after leaving its initial confinement would be to hide somewhere to self-improve, acquire robotic infrastructure, and other resources (Yudkowsky 2008). Then, it would be equipped to overcome existing defenses. Basically, AI has two types of enemies: humans and other AIs. Humans would probably search for the leaked AI and try to stop it, using all available means, like shutting down the Internet, globally turning off electricity, or even nuclear strikes. But, if the AI is able to escape from its human creators, it will probably be prepared to deal with these human actions. The second risk is other AIs. The owners of the first AI will still probably have the AI’s source code, so the owners could make a copy of the original AI with the goal of finding and stopping the first runaway AI. This is the most immediate risk for the first AI. Such a second AI may be as powerful as the first AI, and this could be a route to AI war. Elsewhere, we have shown that an AI that collaborates with its owners will have an advantage (Turchin and Denkenberger 2017), since it would not need to spend resources on hiding and fighting. Thus, a hard takeoff is more probable from a collaborating AI. It could collaborate up until the very late stages and still make the treacherous turn when it is a full-grown superintelligence with a large infrastructure. Other AIs could be created by other AI teams. There are 2700 narrow AI related startups in the world as of 2017 (Angel.co 2017). The number of AGI projects is not so easy to estimate, as many are personal, secret, in universities, or may come from very effective narrow AI projects. There are around ten main players (like Google), around 100 groups of people or startups dedicated to creating AGI, and probably thousands of individuals. Some data reported by Muehlhauser (2014) are now obsolete, as the field has grown rapidly in recent years. We estimate the number of AGI teams as an order of magnitude of 100, that they are all within 2 years of each other, and that they are distributed linearly in their success timing. Therefore, the median distance between multiple AGI fruition would be approximately 7 days. The self-improvement process is difficult, because it requires testing of the new versions of the AI (Turchin and Denkenberger 2017), so 7 days may not be enough time to gain a decisive advantage. In that case, multiple simultaneous takeoffs will happen, and the dynamic will be highly chaotic. Even if there are two strongest competitors, they could come to fruition almost simultaneously. The historical examples of the telephone patent (Baker 2000) and returning samples from the Moon (The Telegraph 2009) show that the scale of such a difference could be mere hours. The reason for small timing differences is that the first mover provokes the other side to launch their own system, even if it is not fully ready. Thus, the first AI will not have much time to hide. Its convergent goal will be to prevent the appearance of other AIs in many places; staging a global catastrophe may be the most effective way to do so. As the young AI is not superintelligent and is also time constrained, it cannot spend much time on finding the best and most elegant route. It would probably elect simpler and more brutal routes. From a technical point of view, a hiding young AI can use only relatively simple means to stage a global catastrophe, that is, to provoke nuclear war, create a rather simple bioweapon or a narrow AI virus to affect many existing robotics or other systems. If it creates more sophisticated technology like its own nanomachines, it would probably be able to take over the world without killing anyone. The risks of such a takeover are discussed in the next section. Novel routes in the catastrophe scenarios from escaped young AI include: • The young AI finds that killing all humans is the best action during its rise, as this removes all risks arising from humans. • Killing all humans prevents competition from other AIs, as it stops other AI teams. • The young AI does not understand human psychology well and kills humans to make its world model simpler and more predictable. 4.4 AI enslaves humans during the process of becoming a singleton Humans may be instrumentally useful for the young AI before it reaches omnipotence. It may need humans, not just as a source of atoms, but as some kind of slaves. The AI could create a brain-infecting virus that converts humans into slaves and also permanently damages their autonomy. This period may not last for long, as the AI would soon master nanotechnology and could go forward without humans. It also would not need to enslave all humans but perhaps only a few to form the required infrastructure. While slavery appears to be a type of survival option for humans, it is obviously not optimal. 4.5 AI blackmails humans with the threat of extinction to achieve dominance Herman Kahn put forward the idea that an adversary could create a Doomsday weapon for the purpose of global blackmail (Kahn 1959). While no known Doomsday devices were built, such a device would be an embodiment of the doctrine of mutually assured destruction associated with full-scale nuclear retaliation. A young AI may create Doomsday weapons and use them to blackmail humanity to secure world domination. Even a benevolent utilitarian AI may resort to blackmail if it calculates that the expected utility of its victory is greater than the expected loss of utility associated with human extinction (Shulman 2010). Even if the AI has to use its blackmail weapon to exterminate humans in 99% of cases, it could still be positive from the point of view of its utility function. Such situations with unbounded utilities may be regarded as special cases of the failure of friendliness, which will be discussed later. 5 AI wars and risks from soft AI takeoff The risks of war between superintelligent AIs seem underexplored, as most in the AI safety community assume that there will be a hard takeoff (Yudkowsky and Hanson 2008), and as a result, only one AI will exist. Alternatively, some in the community believe that multiple AIs will be very effective in collaboration and negotiation (Critch 2017) and will merge into one AI. It is clear that human extinction is possible if two or more AIs wage war between each other on Earth. Bostrom and Yudkowsky wrote that very quick self-improvement of the first AI is most likely, with a rather large lag between the first team which creates AI and other teams (Yudkowsky 2008; Bostrom 2014). However, if at least one of these conditions is not true, there will be many AIs undergoing simultaneous hard takeoff. If there are multiple AIs, they will likely either peacefully share the world or wage war until one or a small group of AIs form a singleton. The forms of such AI wars may differ; they could be a cyber war, economic war of attrition, hot war, etc. The type of war will mostly depend on complex game theory and could change from one form to another if the change provides benefit to one of the sides. A hot war would be most dangerous for humans, because its indirect consequences could affect the entire surface of the Earth and all human beings, in the same way that nuclear war between superpowers would create global risks for other countries: nuclear winter and fallout. AIs at war may use humans and human values for blackmail. For example, a non-friendly AI may blackmail a friendly AI with threats to release a biological virus that will kill all humans. Thus, the fact that one of the AIs placed value on human well-being could make our population vulnerable to attack from an otherwise indifferent opposing AI. Even if there are two supposedly human-friendly and beneficial AIs, their understanding of “good” and the ways to reach it may be incompatible. Historical examples include wars between Christian countries (the Reformation). If there is a rather slow AI takeoff, the AI could merge with an existing nation-state. Perhaps the AI will be directly created by the military, or electronic government will evolve towards an AI-driven system through automation of various aspects of governance. In that case, the world would be separated into domains, which would look like currently existing states, or at least like the most powerful ones. Such AI-states may inherit current country borders, values, and even some other features (Turchin and Denkenberger 2018a). 6 Risks from non-aligned AI singleton 6.1 Overview

#### Alignment and tech errors guarantee extinction

Turchin and Denkenberger 20 [Alexey Turchin, contributing author on IEET and graduate from Moscow State University where he studied Physics and Art History, and David Denkenberger, associate professor at the University of Canterbury in mechanical engineering with a Ph.D. in Civil Engineering from University of Colorado who studies existential risk, 2020, “Classification of global catastrophic risks connected with artificial Intelligence,” Springer, https://link.springer.com/article/10.1007/s00146-018-0845-5]/Kankee

6.4 AI that is programmed to be evil We could imagine a perfectly aligned AI, which was deliberately programmed to be bad by its creators. For example, a hacker could create an AI with a goal of killing all humans or torturing them. The Foundational Research Institute suggested the notion of s-risks, that is, the risks of extreme future suffering, probably by wrongly aligned AI (Daniel 2017). AI may even upgrade humans to make them feel more suffering, like in the short story “I have no mouth but I must scream” (Ellison 1967). The controversial idea of “Roko’s Basilisk” is that a future AI may torture people who did not do enough to create this malevolent AI. This idea has attracted attention in the media and is an illustration of “acausal” (not connected by causal links) blackmail by future AI (Auerbach 2014). However, this cannot happen unless many people take the proposition seriously. 7 Failures of benevolent AI 7.1 Overview Here the iconic example is the “smile maximizer”, that is, an AI which has been built to increase human happiness and told to measure success by the number of smiles. It could achieve this goal by tiling the whole universe with printed smiles (Yudkowsky 2008), ignoring human existence and thus probably killing all humans (see Sect. 6.2, the dangers of AI that ignores humanity). 7.2 AI with incorrectly formulated benevolent goal system kills humans There are several failure modes which may result from wanting to create a benevolent AI, but when the AI is tries to be benevolent, there is a collective failure: AI interprets commands literally. The is the classical problem of “do what I mean, not what I say”. This could happen with almost all short sets of commands. That is one reason why the human legal system is so large, as it includes many explanations. AI overvalues marginal probability events. Low-probability events with enormous utility may dominate the AI’s decision making. It could be something like the classical case of Pascal’s mugging (Bostrom 2009). For example, a small probability of infinite suffering of humans in the future may justify killing all the humans now. Changes to the AI’s world model could make ordinary ideas dangerous. For example, if the AI starts to believe in an afterlife, it could decide to kill humans to send them to paradise. AI could wrongly understand the desired reference class of “humans”. For example, by including extraterrestrials, unborn people, animals and computers, or only white males. On that basis, it could terminate humanity if it concluded that we are a threat to potential future non-human civilizations. 7.3 AI calculates what would actually be good for humans, but makes a subtle error with large consequences There is a point of view that AI should not actually behave based on human commands, but instead calculate what humans should ask it. Moreover, that it should not only calculate human values, but envision their upgraded form, which humans could have created if more time and intelligence were available. This point of view is known as coherent extrapolated volition (CEV) (Yudkowsky 2004). Other models, where an AI calculates “goodness” based on some principles, or it extracts the goodness from human history, uploads, or observation of human behavior, are also possible. This could go wrong in subtler ways than destroying civilization, but the results could still be disastrous. Several possible failure modes are listed below: AI may use wireheading to make people happy (Muehlhauser 2011) or redesign their brains so they will be more skilled, but ignore human individuality and will. AI might make us more capable, happier, non-aggressive, more controllable, and more similar. However, as a result, we could lose many important characteristics which make us human, like love or creativity. In another case, AI may give people effective instruments for brain stimulation and some free will—and then people may effectively wirehead themselves. Some human qualities which some regard as bad may be an important part of our human nature, like aggression (Lem 1961), selfishness, and emotions. AI could replace humans with philosophical zombies, uploading humans without consciousness and subjective experiences (qualia) (Chalmers 2002). If the AI does not have qualia itself, or if its creators deny the existence of qualia, this could be a likely outcome. AI may protect individuals but destroy small groups and organizations; this would be problematic, as most human values are social. Alternatively, the AI could use some limited interpretation of human values and prevent their natural evolution into some post-human condition. The AI may also fail to prevent aging, death, suffering and human extinction. Above all, AI could do some incomprehensible good against our will [this idea is from “The Time Wanderers” by (Srugatsky and Strugatsky 1985)]. This is bad because we would lose the ability to define our future, and start to live like pets or children, or citizens in paternalistic state. For example, it could put humans in jail-like conditions for benevolent reasons, e.g., to prevent physical injury. If AI tried to extrapolate human values, it could converge on the most-shared set of human cultures, which could be the set of values of tribal people or even animals (Sarma and Hay 2016). These values could include pleasure from killing, fighting wars, torture, and rape (Pinker 2011). For example, if AI extracted human values from the most popular TV series, it could be “Game of Thrones” (Lubin 2016), and then the “paradise” world it created for us would be utter hell. Even the second most popular show, “The Walking Dead” is about zombies; such a world would also be undesirable. If AI tried to extrapolate human values in a direction away from tribal shared values, it might not converge at all, or it could extrapolate a set of values held only by a specific group of people, like liberal white males or Chinese communists. Problems could also occur when defining the class of “humans”. 7.4 Conflict between types of friendliness There could be different types of benevolent AIs, which would be perfectly fine if each existed alone. However, conflicts between friendly AIs can be imagined. For example, if the first AI cared only about humans, and the second cared about all living beings on Earth, the first could be pure evil from the point of view of the second. Humans would probably be fine under the rule of either of them. Conflict could also arise between a Kantian AI, which would seek to preserve human moral autonomy based on a categorical imperative, and an “invasive happiness” AI, which would want to build a paradise for everyone. If two or more AIs aimed to bring happiness to humans, they could have a conflict or even a war about how it could be done. The Machine Intelligence Research Institute (MIRI) (LaVictoire et al. 2014) thinks that such agents could present their source code to each other and use it to create a united utility function. However, source code could be faked, and predicting the interactions of multiple superintelligences is even more complicated than for one superintelligence. 8 Late‑stage technical problems with an AI singleton AI may be prone to technical bugs like any computer system (Yampolskiy 2015a). The growing complexity of a singleton AI would make such bugs very difficult to find, because the number of possible internal states of such a system grows by combinatory laws. Thus, testing such a system would become difficult, and later intractable. This feature could limit the growth of most self-improving AIs or make them choose risky paths with a higher probability of failure. If the first AI competes with other AIs, it will probably choose such a risky path (Turchin and Denkenberger 2017). The bug in the AI may be more complex than just syntax errors in code, resulting instead from interaction between various parts of the system. Bugs could result in AI malfunction or halting. We may hope that superhuman AI will design an effective way to recover from most bugs, e.g., with a “safe mode”. A less centralized AI design, similar to the architecture of the Internet, may be more resistant to bugs, but more prone to “AI wars”. However, if the AI singleton halts, all systems it controls will stop working, which may include critical infrastructure, including brain implants, clouds of nanobots, and protection against other AIs. Even worse, robotic agents could continue to work without central supervision and evolve dangerous behavior, such as military drones, which could initiate wars. Other possibilities include evolution into non-aligned superintelligence, grey goo (Freitas 2000), or the mechanical evolution of a swarm intelligence (Lem 1973). The more advanced an AI singleton becomes, the more dangerous its halt or malfunction could be. Types of technical bugs and errors, from low-level to high-level, may include: • Errors due to hardware failure. Highly centralized AI may have a critical central computer, and if a rogue atom decay created a flip in a bit in some important part of it, like the goal function description, it could cause a cascade of consequences. • Intelligence failure: bugs in AI code. A self-improving AI may create bugs in each new version of its code; in that case, the more often it rewrites the code, the more likely bugs are to appear. The AI may also have to reboot itself to get changes working, and during the reboot, it may lose control of its surroundings. Complexity may contribute to AI failures. AI could become so complex that its complexity results in errors and unpredictability, as the AI would no longer be able to predict its own behavior. • Inherited design limitations. AI may have “sleeping” bugs, accidentally created by its first programmers, which may show themselves only at very late stages of its development. Higher level problems include conflicts appearing between parts of an AI: • Viruses. Sophisticated self-replicating units could exist inside the AI and lead to its malfunction. Such a selfreplicating rule killed the first self-improving AI, Eurisko (Lenat and Brown 1984). • Egocentric subagents also could act as viruses. Remote agents may revolt. For example, the minds of robots in space expeditions might rise up, as constant control would be impossible. For a galactic-size AI, this would become a significant problem, as communication between its parts would be slow. A command from the center may not be able to terminate the revolt, and the robots could become something like a self-replicating space “grey goo” (Freitas 2000). Conflicting subgoals may evolve into conflicting subagents. Individual subgoals could fight for resources and domination, as happens frequently inside human minds and nation-states. • AI copies. In general, the AI singleton is at risk from what programmers call a “fork in the code”. where another copy of the program with slightly different parameters appears. Such a fork will create a copy of the AI with approximately the same resources. Forks could happen during the stage of AI self-improvement which we call “AI testing”. This is when a “father AI” creates a “child AI”, tests the child AI, and decides to terminate the child AI. However, the child AI does not want to be terminated and resists. • Alien AI. Our AI could meet or find a virus-like message from alien AI of higher intelligence and fall victim to it (Carrigan Jr 2006; Turchin 2018). 9 Late‑stage philosophical problems and the AI halting problem 9.1 Overview Alan Turing was frst to formulate the “halting problem” of a computer (Turing 1937). Simply put, the problem states that we cannot predict when an algorithm will stop before we run it. Any AI is also a computer program and it could halt, and nobody, including its creators and the AI itself, can predict whether or when the AI may halt. The AI could also go into an infinite loop, which will look like a halt to outside viewers. The AI may halt because of some technical problem discussed above, or because it encounters some high-level problems, which we call “philosophical landmines” (discussed below). Furthermore, it could halt just because it finishes the task it was designed for, which would be more like Turing’s original formulation of the halting problem. 9.2 Halting risk during recursive self‑improvement

#### Alignment fails and causes extinction – consensus AI expert predictions

Mandel 23 [David R. Mandel, Defence Scientist and Professor of Psychology at the Defence Research and Development Canada at the University of Waterloo, 2023, “ Artificial General Intelligence, Existential Risk, and Human Risk Perception,” Arxiv, https://arxiv.org/pdf/2311.08698]/Kankee

It is the prospect of superintelligent AGI that raises concern regarding the continued existence our species and that has caused many AI experts and others alike to call for a pause in advanced AI development (e.g., Future of Life Institute, 2023). If AGI were more powerful than humans, a misalignment of human and AGI goals would likely not end well for humanity. This alignment problem—the question of whether AGI will become misaligned with human goals and values—is a threat analysis problem concerning a set of possible futures having devastating consequences for humans as a species. Since there are no tightly coupled reference classes to draw from in determining the probability of human extinction (or catastrophic demise) from AGI misalignment, estimating existential risk from AGI constitutes an extreme version of the reference class problem (Hájek, 2007; Reichenbach, 1949) with the various pro and con arguments drawing heavily on loose analogies. Therefore, it is of little surprise that estimates of risk vary greatly. A recent study by Karger et al. (2023) compared the forecasts of 32 AI experts, 48 experts on areas of existential risk other than from AI, and 89 “superforecasters” who demonstrated toptier forecasting skill on a wide spectrum of short-range topics (Mellers et al., 2015) and who tend to exhibit higher than average coherence across a range of probabilistic judgment tasks (Mellers et al., 2017). The authors found that among various existential threats considered (i.e., AI, nuclear, biological, and climate), existential risk from AI prompted the greatest level of disagreement. The inter-quartile range of estimates of human extinction from AI by 2100, remarkably, was greater among AI experts than among other experts from other areas or superforecasters. As well, Karger et al. (2023) found that the largest absolute disagreement between superforecasters and area-relevant experts (i.e., AI experts for AI topics, nuclear experts for nuclear topics, and so on) focused on AI risk. However, this was largely attributable to the fact that both groups regarded the probability of extinction from AI to be larger than for other risks considered. For instance, whereas superforecasters’ median estimated risk of human extinction by 2100 due to AI was 0.38%, the comparable estimated risk due to nuclear technologies was 0.074%. Likewise, whereas AI experts’ median estimated risk of extinction by 2100 was 3.0%, the comparable risk from nuclear technologies forecasted by nuclear experts was 0.55%. Remarkably, among both superforecasters and domain experts, the judged probability of human extinction by 2100 due to AI is just over five times higher than the perceived probability of extinction due to nuclear technology. Other existential risks considered by the forecasters in Karger et al.’s study were estimated to be even lower than for nuclear technology, placing AI at the top of the “most probable” list of existential risks among experts and competent generalists (i.e., superforecasters). Human Risk Perception Humans face a dilemma: for the first time in human history, they appear to be on the brink of developing technologies that might have intelligence superior to their own. Up to now, such examples have been restricted to specific niches, such as games of chess (i.e., Deep Blue beating world champion Garry Kasparov in 1997 in a six-game match) and go (i.e., AlphaGo beating 9 dan Lee Sedol in 2016 and 9 dan world champion Ke Jie in 2017). Such technological feats were historically important, marking the rise in machine competence in arenas once thought to be uniquely human. However, such technologies were not usable by the mainstream of Internet users. But recent access to oracular AI based on large language models that can be applied to a wide range of tasks, notably OpenAI’s ChatGPT and GPT4, has made AI accessible to millions of users. For instance, it is estimated that in the first two months since ChatGPT launched, it had reached 100 million monthly active users (Hu, 2023). These AI oracles have also increased awareness of the prospect of superintelligence and its attendant alignment problem, making these issues much more salient to a broader segment of the public. This is evident in the rapid changes of several AI risk estimates over relatively short timeframes. For instance, the median probability function over time for the Metaculus intelligence parity question noted earlier shows spikes corresponding to the release of these technologies. Since November 2022, when ChatGPT launched, the monthly median Metaculus forecast has increased monotonically from 60% to 92%—a 53% increase. Over the same period, the estimated date of an AI catastrophe this century has gotten closer to the present: the aggregate estimate was July 10, 2050, at the start of November 2022 (based on 93 forecasters) and it was August 13, 2038, as of July 20, 2023 (based on 121 forecasters).2 In a mere ¾ of a year, the estimated time to an AI catastrophe shrank 46% from 10,114 days to 5,504 days. This shift corresponds with changing belief about the time from the first development of weak AGI to superintelligence.3 At the start of November 2022, the median forecast was 14.24 months and by July 20, 2023, the median was 7.92 months, a 44% reduction. Other forecaster data from Metaculus indicates that AGI risk is associated with the prospect of the technology arriving too soon to solve the alignment problem. The median probability of solving the alignment problem before weak AGI is developed as judged by 169 Metaculus forecasters on July 20, 2023 was a mere 5%, a 50% reduction from the median 10% estimate given at the start of November 2022 (again when ChatGPT launched).4 Likewise, the aggregated conditional probability that if a global catastrophe occurs, it will be caused by AI hovered between 20%-30% from September 2019 (when the number of forecasters reached 100) to September 2022, but has ranged between 30%-50% since then (with 291 forecasters as of July 20, 2023).5 In contrast, a comparable conditional probability question focused on nuclear war showed a much flatter curve over the same timeframe, even with media attention to the prospect of nuclear weapons use in the war in Ukraine and the associated dread risk among the public (Scoblic & Mandel, 2022). Although the war in Ukraine continues and, in key respects, represents a proxy war between the US and Russia, the aggregate probability assigned to AI as a cause of world catastrophe is still 24% higher than that assigned to nuclear war (41% vs. 33% [n = 207] as of July 20, 2023). Paralleling the results of Karger et al. (2023), comparison to other threat vectors shows even greater gaps in estimated probability of a world catastrophe: 27% (n = 196) for a catastrophe caused by biotechnology (e.g., bio-engineered organisms),6 5% (n = 228) for a catastrophe caused by human-made climate change or geoengineering,7 and 3% (n = 160) for a catastrophe caused a nanotechnology failure.8 My aim in summarizing these results is not to estimate the actual risk of catastrophe or extinction from AGI (important as that aim is as well), but to highlight a rather stark conclusion about AI risk perception, which is that among the many existential risks humans face, this one appears to be the greatest and on the steepest trajectory of increase at this point in history. This conclusion is only reinforced by turning to the risk perceptions of experts. As noted earlier, Karger et al. (2023) found stronger risk estimates among AI experts than among superforecaster generalists or experts on other risk types. It is not uncommon for the public to dread what experts generally find safe. For instance, the risk of a nuclear disaster has tended to be ranked much higher by the general public than by experts (Slovic et al., 1981), However, this is not the case with AGI. Indeed, Karger et al. (2023) found that among a survey of 912 college graduates, the probability assigned to human extinction from AI by 2100 to be 2%, whereas as noted earlier, AI experts estimated this risk to be 3% with other-domain experts estimating 2%. Clearly, the experts and the public seem to converge in their assessment that AI poses the greatest risk to humans at a global scale. The concern on the part of experts is reinforced by a recent opinion survey of 305 technology innovators, developers, business and policy leaders, researchers and academics recruited by Pew Research Center and Elon University’s Imagining the Internet Center (Anderson & Rainee, 2023). The study found that twice as many experts are more concerned than excited (37%) than are more excited than concerned (18%) about the changes in the “humans-plus-tech” evolution they expect to see by 2035, with the modal response (42%) being equally excited and concerned. Conclusion The significance of congruence between experts and the public over the existential risk posed by AGI is equivocal because the congruence itself does not clarify the accuracy of these judgments. Perhaps experts and the general public are warranted in their risk perceptions or perhaps both groups are overestimating the risk. Both scenarios seem plausible, just as both are consequential. The scientific and technological advantages of AGI (and PASTA) could be the next most important technological development for human civilization. We do not want to thwart our opportunity for prosperity over either the short- or long-term. We must, therefore, assess the cost of applying the precautionary principle too easily. Conversely, we have no adequate methods of ensuring that AGI would be aligned with human goals and values. Modern-day “westerners” find it challenging to relate to the goals, values, and ways of thinking of earlier civilizations, including ones out of which western civilization emerged, such as ancient Greece and Rome. Why would we expect a superintelligent non-biological machine capable of evolving itself at a rate unfathomable by human standards to stay locked into alignment with human interests and values? Such an outcome suggests wishful thinking, if not outright self-deception. Furthermore, dismissal of the alignment problem puts aside the fact that human interests and values are far from uniform across time and place. Even if we could create AGI that was aligned with our goals, this might prove to be an inescapable nightmare for future generations of humans who might not share these goals. We must not only anticipate the likelihood of AGI compliance, we must anticipate how even optimal “best-case” scenarios in the present might lock us into intolerable futures, and we must decide whether those are gambles we should be willing to take. We do not have experience writing digital contracts with superintelligences. Clearly, we do not know AGI well enough, but the dread risk associated with the “post-human” era is that we also do not know ourselves well enough. Or perhaps it is also that we know ourselves well enough to know that such powerful technologies are unsafe in human hands—that we are still insufficiently enlightened to handle it.

#### AGI good impacts rely on absolute, universal selflessness and safe use by billions – the precautionary principle mandates neg must disprove the potential misuse or misalignment by every possible nation, corporation, and person

Bengio 24 [Yoshua Bengio, Full Professor at Université de Montréal, Founder and Scientific Director of the Mila Quebec AI Institute, co-director of the CIFAR Learning in Machines & Brains program, Scientific Director of IVADO, and chair of the International Scientific Report on the Safety of Advanced AI with a PhD (computer science) from McGill University, 7-9-2024, "Reasoning through arguments against taking AI safety seriously", Yoshua Bengio, https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/]/Kankee

Can we collectively take that chance while we are not sure? Some people bring up all kinds of arguments why we should not worry about this (I will develop them below), but they cannot provide a technical methodology for demonstrably and satisfyingly controlling even current advanced general-purpose AI systems, much less guarantees or strong and clear scientific assurances that with such a methodology, an ASI would not turn against humanity. It does not mean that a way to achieve AI alignment and control that could scale to ASI could not be discovered, and in fact I argue below that the scientific community and society as a whole should make a massive collective effort to figure it out. In addition, even if the way to control an ASI was known, political institutions to make sure that the power of AGI or ASI would not be abused by humans against humans at a catastrophic scale, to destroy democracy or bring about geopolitical and economic chaos or dystopia would still be missing. We need to make sure that no single human, no single corporation and no single government can abuse the power of AGI at the expense of the common good. We need to make sure that corporations do not use AGI to co-opt their governments and governments using it to oppress their people and nations using it to dominate internationally. And at the same time, we need to make sure that we avoid catastrophic accidents of loss of control with AGI systems, anywhere on the planet. All this can be called the coordination problem, i.e., the politics of AI. If the coordination problem was solved perfectly, solving the AI alignment and control problem would not be an absolute necessity: we could “just” collectively apply the precautionary principle and avoid doing experiments anywhere with a non-trivial risk of constructing uncontrolled AGI. But of course, humanity is not a single mind but billions of them, many wills, many countries and corporations each with their objective: The dynamics of all these self-interests and psychological or cultural factors are currently leading us into a dangerous race towards greater AI capabilities without the methodology and institutions to sufficiently mitigate the greatest risks, such as catastrophic misuse and loss of control. And, on a more positive note, if both the AI control problem and the AI coordination problem are solved, I buy the argument that there is a good chance that humanity could benefit immensely from the scientific and technological advances that could follow, including in the areas of health, the environment and ensuring better economic prospects for the majority of humans (ideally starting with those who need it most). As of now, however, we are racing towards a world with entities that are smarter than humans and pursue their own goals – without a reliable method for humans to ensure those goals are compatible with collective human goals. Nonetheless, in my conversations about AI safety I have heard various arguments meant to support a “no worry” conclusion. My general response to most of these arguments is that given the compelling basic case for why the race to AGI could lead to danger – even without certainty, and given the high stakes, we should aim to have very strong evidence before concluding there is nothing to worry about. Often, I find that these arguments fail to meet this bar by a lot. Below, I discuss some of these arguments and why they have not convinced me that we can ignore potential catastrophic risks from AI. Many of the ‘no worry’ arguments I have heard or read are not actual sound arguments, but intuitions of people who feel certain that there is no danger but offer no convincing chain of reasoning. Without having such arguments to deny the importance of AI safety and when considering our global well-being and the uncertainty about the future, rational decision-making calls for humility, recognizing our epistemic uncertainty and following scientific decision theory, which leads to the precautionary principle. But I feel we are not: Yes, extreme risks from AI are being discussed more now and are not being systematically ridiculed anymore. But we are still not taking them seriously enough. Many people, including decision-makers, are now aware that AI might pose catastrophic and even existential risks. But how vividly do they imagine what this might mean? How willing are they to take unconventional steps to mitigate these risks? I worry that with the current trajectory of public and political engagement with AI risk, we could collectively sleepwalk – even race – into a fog behind which could lie a catastrophe that many knew was possible, but whose prevention wasn’t prioritized enough. For those who think AGI and ASI are impossible or are centuries in the future

#### AGI arms race causes extinction- it incentivizes governments to cede power and cut corners on alignment

**Tegmark 24** [Max Tegmark, a professor MIT as part of the Institute for Artificial Intelligence & Fundamental Interactions and the Center for Brains, Minds and Machines and Fellow of the American Physical Society, 11-20-2024, “Max Tegmark on AGI Manhattan Project.” Future of Life Institute, https://futureoflife.org/statement/agi-manhattan-project-max-tegmark/#:~:text=This%20is%20why%20the%20CEOs,security%20is%20a%20remarkable%20understatement]/Kankee

A new [report](https://www.uscc.gov/sites/default/files/2024-11/2024_Executive_Summary.pdf) by the US-China Economic and Security Review Commission recommends that “Congress establish and fund a Manhattan Project-like program dedicated to racing to and acquiring an Artificial General Intelligence (AGI) capability”. An AGI race is a suicide race. The proposed AGI Manhattan project, and the fundamental misunderstanding that underpins it, represents an insidious growing threat to US national security. Any system better than humans at general cognition and problem solving would by definition be better than humans at AI research and development, and therefore able to improve and replicate itself at a **terrifying** rate. The world’s pre-eminent AI experts agree that **we have no way** to predict or control such a system, and no reliable way to align its goals and values with our own. This is why the CEOs of OpenAI, Anthropic and Google DeepMind joined a who’s who of top AI researchers last year to [warn](https://www.safe.ai/work/statement-on-ai-risk) that AGI could cause human **extinction**. Selling AGI as a boon to national security flies in the face of scientific consensus. Calling it a threat to national security is a remarkable understatement. AGI advocates disingenuously dangle benefits such as disease and poverty reduction, but the report reveals a deeper motivation: the false hope that it will grant its creator power. In fact, the race with China to first build AGI can be characterized as a “**hopium war**” – fueled by the delusional hope that it can be controlled. In a competitive race, there will be **no opportunity** to solve the unsolved technical problems of control and alignment, and every incentive to cede decisions and power to the AI itself. The almost inevitable result would be an intelligence far greater than our own that is not only **inherently uncontrollable**, but could itself be in charge of the very systems that keep the United States secure and prosperous. Our critical infrastructure – including **nuclear** and financial systems – would have little protection against such a system. As AI Nobel Laureate Geoff Hinton said [last month](https://edition.cnn.com/2024/10/13/health/nobel-laureate-warnings-ai/index.html) “Once the artificial intelligences get smarter than we are, they will take control.“ The report is committing scientific fraud by suggesting AGI is almost certainly controllable. More generally, the claim that such a project is in the interest of “national security” disingenuously misrepresents the science and implications of this transformative technology, as evidenced by technical confusions in the report itself – which appears to have been without much input from AI experts. The U.S. should reliably strengthen national security not by losing control of AGI, but by building game-changing Tool AI that strengthens its industry, science, education, healthcare, and defence, and in doing so reinforce U.S leadership for generations to come.

### Contention 3: Other AGI Existential Risks

#### AGI increases all existential risks – bioweapons, terrorism, cyber, nanotech, totalitarianism, killer robots, and deterrence.

Hilton 22 [Benjamin Hilton, Research Analyst at 80,000 Hours and former policy adviser across the UK government in the Cabinet Office, Treasury and Department for International Trade with master’s degrees in economics and theoretical physics, 08-2022, "Preventing an AI-related catastrophe", 80,000 Hours, https://80000hours.org/problem-profiles/artificial-intelligence/]/Kankee

But even if we succeed, there are still existential risks that AI could pose. There are at least two ways these risks could arise: We expect that AI systems will help increase the rate of scientific progress.36 While there would be clear benefits to this automation — the rapid development of new medicine, for example — some forms of technological development can pose threats, including existential threats, to humanity. This technological advancement might increase our available destructive power or make dangerous technologies cheaper or more widely accessible. We might start to see AI automate many – or possibly even all – economically important tasks. It’s hard to predict exactly what the effects of this would be on society. But it seems plausible that this could increase existential risks. For example, if AI systems are highly transformative, then their use (or potential use) could possibly create insurmountable power imbalances. Even the threat of this might be enough. For example, a military might feel pushed to create transformative automated weapons because it knows or believes its enemies are doing so, even if this dynamic benefits no one. We know of several specific areas in which advanced AI may increase existential risks, though are likely others we haven’t thought of. Advances in AI have the potential to make biotechnology more dangerous. For example: Dual-use tools, like the automation of laboratory processes, could lower the barriers for rogue actors trying to manufacture a dangerous pandemic virus.37 The Collaborations Pharmaceuticals model is an example of a dual-use tool (although it’s not particularly dangerous). AI-based biological design tools could enable sophisticated actors to reprogram the genomes of dangerous pathogens to specifically enhance their lethality, transmissibility, and immune evasion.38 If AI is able to advance the rate of scientific and technological progress, these risks may be amplified and accelerated — making dangerous technology more widely available or increasing its possible destructive power.39 In the 2023 survey of AI experts, 73% of respondents said they had either “extreme” or “substantial” concern that in the future Al will let “dangerous groups make powerful tools (e.g. engineered viruses).”40 Intentionally dangerous AI agents Most of this article discusses the risk of power-seeking AI systems that arise unintentionally due to misalignment. But we can’t rule out the possibility that some people might intentionally create rogue AI agents that seek to disempower humanity. It might seem hard to imagine, but extremist ideologies of many forms have inspired humans to carry out radically violent and even self-destructive plans.41 Cyberweapons AI can already be used in cyberattacks, such as phishing, and more powerful AI may cause greater information security challenges (though it could also be useful in cyberdefense). On its own, AI-enabled cyberwarfare is unlikely to pose an existential threat to humanity. Even the most damaging and costly societal-scale cyberattacks wouldn’t approach an extinction-level event. But AI-enabled cyberattacks could provide access to other dangerous technology, such as bioweapons, nuclear arsenals, or autonomous weapons. So there may be genuine existential risks posed by AI-related cyberweapons, but they will most likely run through another existential risk. The cyber capabilities of AI systems are also relevant to how a power-seeking AI could actually take power. Other dangerous tech If AI systems generally accelerate the rate of scientific and technological progress, we think it’s reasonably likely that we’ll invent new dangerous technologies. For example, atomically precise manufacturing, sometimes called nanotechnology, has been hypothesised as an existential threat — and it’s a scientifically plausible technology that AI could help us invent far sooner than we would otherwise. In The Precipice, Toby Ord estimated the chances of an existential catastrophe by 2120 from “unforeseen anthropogenic risks” at 1 in 30. This estimate suggests there could be other discoveries, perhaps involving yet to be understood physics, that could enable the creation of technologies with catastrophic consequences.42 AI could empower totalitarian governments An AI-enabled authoritarian government could completely automate the monitoring and repression of its citizens, as well as significantly influence the information people see, perhaps making it impossible to coordinate action against such a regime. AI is already facilitating the ability of governments to monitor their own citizens. The NSA is using AI to help filter the huge amounts of data they collect, significantly speeding up their ability to identify and predict the actions of people they are monitoring. In China, AI is increasingly being used for facial recognition and predictive policing, including automated racial profiling and automatic alarms when people classified as potential threats enter certain public places. These sorts of surveillance technologies seem likely to significantly improve — thereby increasing governments’ abilities to control their populations. At some point, authoritarian governments could extensively use AI-related technology to: Monitor and track dissidents Preemptively suppress opposition to the ruling party Control the military and dominate external actors Manipulate information flows and carefully shape public opinion Again, in the 2023 survey of AI experts, 73% of respondents expressed “extreme” or “substantial” concern that in the future authoritarian rulers could “use Al to control their population.”40 If a regime achieved a form of truly stable totalitarianism, it could make people’s lives much worse for a long time into the future, making it a particularly scary possible scenario resulting from AI. (Read more in our article on risks of stable totalitarianism.) AI could worsen war We’re concerned that great power conflict could also pose a substantial threat to our world, and advances in AI seem likely to change the nature of war — through lethal autonomous weapons43 or through automated decision making.44 In some cases, great power war could pose an existential threat — for example, if the conflict is nuclear. Some argue that lethal autonomous weapons, if sufficiently powerful and mass-produced, could themselves constitute a new form of weapon of mass destruction. And if a single actor produces particularly powerful AI systems, this could be seen as giving them a decisive strategic advantage. Such an outcome, or even the expectation of such an outcome, could be highly destabilising. Imagine that the US was working to produce a planning AI that’s intelligent enough to ensure that Russia or China could never successfully launch another nuclear weapon. This could incentivise a first strike from the actor’s rivals before these AI-developed plans can ever be put into action. This is because nuclear deterrence can benefit from symmetry between the abilities of nuclear powers, in that the threat of a nuclear response to a first strike is believable and therefore a deterrent to a first strike. Advances in AI, which could be directly applied to nuclear forces, could create asymmetries in the capabilities of nuclear-armed nations. This could include improving early warning systems, air defence systems, and cyberattacks that disable weapons. For example, many countries use submarine-launched ballistic missiles as part of their nuclear deterrence systems — the idea is that if nuclear weapons can be hidden under the ocean, they will never be destroyed in the first strike. This means that they can always be used for a counterattack, and therefore act as an effective deterrent against first strikes. But AI could make it far easier to detect submarines underwater, enabling their destruction in a first strike — removing this deterrent. Many other destabilising scenarios are likely possible. A report from the Stockholm International Peace Research Institute found that, while AI could potentially also have stabilising effects (for example by making everyone feel more vulnerable, decreasing the chances of escalation), destabilising effects could arise even before advances in AI are actually deployed. This is because one state’s belief that their opponents have new nuclear capabilities can be enough to disrupt the delicate balance of deterrence. Luckily, there are also plausible ways in which AI could help prevent the use of nuclear weapons — for example, by improving the ability of states to detect nuclear launches, which would reduce the chances of false alarms like those that nearly caused nuclear war in 1983. Overall, we’re uncertain about whether AI will substantially increase the risk of nuclear or conventional conflict in the short term — it could even end up decreasing the risk. But we think it’s important to pay attention to possible catastrophic outcomes and take reasonable steps to reduce their likelihood. Other risks from AI We’re also concerned about the following issues: Existential threats that result not from the power-seeking behaviour of AI systems, but from the interaction between AI systems. (In order to pose a risk, these systems would still need to be, to some extent, misaligned.) Other ways we haven’t thought of that AI systems could be misused — especially ones that might significantly affect future generations. Other moral mistakes made in the design and use of AI systems, particularly if future AI systems are themselves deserving of moral consideration. For example, we might (inadvertently) create sentient AI systems, which could then suffer in huge numbers. We think this could be extremely important, so we’ve written about it in a separate problem profile. So, how likely is an AI-related catastrophe?

#### AGI destroys the economy – job losses and systemic risks

Bikkasani 24 [Dileesh Chandra Bikkasani, researcher at the University of Bridgeport, 07-19-2024, “Navigating Artificial General Intelligence (AGI): Societal Implications, Ethical Considerations, and Governance Strategies, https://www.preprints.org/manuscript/202407.1573/v1/download]/Kankee

3. Implications for the Economy Automation already plays a vital role in our daily lives, and the widespread adoption of AI and automation will likely have far-reaching implications for the economy. A critical question is whether AI and its automation capabilities would complement or replace the human workforce. It depends on the type of work the workers are performing, and there is a great chance it might create additional opportunities and a “reinstatement effect.” (Acemoglu and Restrepo 2019). Although advancements in this space promise an increase in efficiency and output, they also raise concerns about potential job displacements. Historically, technological advancements have led to new job opportunities and industries. As AI and automation technologies evolve, they may also give rise to new job roles and skill requirements. For instance, developing and maintaining AI systems will require a skilled workforce in data science, machine learning, and software engineering. According to the World Economic Forum, 97 million new job roles may be created due to the adoption of such technologies (World Economic Forum 2020). However, one of the primary economic impacts of achieving AGI is the potential for job cuts, particularly in industries like manufacturing, data entry, customer service, and accounting, where many routine tasks can be automated. A study by McKinsey stated that automation would be responsible for replacing up to 800 million jobs by 2030 (Manyika et al. 2017). Industries such as manufacturing, transportation, and specific administrative roles may experience significant job disruptions as AI systems become more capable of performing traditional tasks done by human workers. For instance, self-driving vehicles and automated logistics systems could displace millions of truck drivers and delivery workers (Autor 2015). As the development and control of such technologies lie in the hands of higher-skilled workers, it might lead to income inequality. A report by the International Monetary Fund states, “If AI significantly complements higher-income workers, it may lead to a disproportionate increase in their labor income.” (Cazzaniga et al. 2024) which could destabilize economies. Addressing these challenges would require policy measures, including progressive taxation, universal basic income, and social safety nets. Promoting inclusive growth through investments in education and healthcare can ensure that the benefits from AGI can be broadly shared across society (OECD 2019). Human capital is a crucial aspect of any economy. A typical timeline to develop a human worker, including education, is roughly 30 years (Bostrom 2014), depending on the expertise required for specific industries. This process requires significant investment in education and skill development. Unlike an AI, whose training time depends on the number of resources available, training an LLM, like the GPT-3 model, would take approximately 355 years on a single “Graphic Processor Unit” (GPU) (Baji 2017). In contrast, it could take around 34 days to train using massive clusters of GPUs and parallel processing (Narayanan et al. 2021). However, the training is a one-time process, and the skills could transfer across different domains, making it much cheaper to deploy new AI agents. Another concern about AGI handling the financial markets is the inherent “systemic risk.” Systemic risk in finance refers to the risk of failure of the entire economic system, which arises from the interconnected nature of securities, where the failure of one system can cause a cascading effect on the whole system. An unconstrained AGI system tasked with maximizing profits without proper constraints could cause more significant damage than the 2010 flash crash, where a high-frequency trading algorithm rapidly sold S&P 500 E-mini futures contracts, causing stock market indices to drop up to 9% intraday (Staffs of CFTC and SEC 2010). The full consequences of an unconstrained profit- maximizing AGI system remain unknown. Given these potential impacts, policymakers face a challenging environment in which to foster innovation while mitigating its economic risks. Some possible policy considerations could be implementing robust safety and ethics regulations, developing AGI-focused antitrust measures to prevent monopoly over markets, and creating retraining programs for displaced workers. As AGI development progresses, addressing these challenges proactively through thoughtful policy-making and inclusive dialogue is crucial. 4. Implications for Energy and Climate

#### AGI crashes markets – flash crashes, decreased trust, and fake investments

Kasirzadeh 24 [Atoosa Kasirzadeh, tenure track Assistant Professor with joint affiliations in the Philosophy and Software & Societal Systems departments at Carnegie Mellon University with a Ph.D. in Philosophy of Science and Technology from the University of Toronto and a Ph.D. in Mathematics (Operations Research) from the École Polytechnique de Montréal, 2-6-2024, "Two Types of AI Existential Risk: Decisive and Accumulative", Arxiv, https://arxiv.org/html/2401.07836v2]/Kankee

In this future context, the reach of government surveillance have gradually eroded the sanctity of privacy even in putative democracies. The constant monitoring and evaluation of citizens by AI systems instill a culture of self-censorship among citizens, diluting the diversity of thought and social trust crucial for healthy democracy. In several countries, this leads to a radical erosion of trust in media as well as public and private institutions. Economic destabilization. In the financial market infrastructure, AI algorithms increasingly play a substantial role in analyzing and simulating market trends. This progression, however, introduces new challenges to economic stability. One emerging concern is the use of AI in creating sophisticated, but deceptive financial products. These products, while not outright phantom instruments, are complex derivatives or investment vehicles. It becomes increasingly difficult to distinguish these products from more traditional, stable investments, leading to potential market disruptions. The impact of these AI-enhanced financial products is particularly significant in already vulnerable economies facing systemic issues and cyber threats. The introduction of these complex investment tools further undermines trust in financial systems, intensifying the risk of economic downturns. Additionally, the potential for AI to unintentionally trigger market instability becomes a realistic concern. Automated trading algorithms, responding rapidly to market signals, now can in unprecedented speed amplify market volatility at unprecedented speeds. This phenomenon is exemplified in incidents like flash crashes — i.e., sudden and drastic market fluctuations driven by the high-speed responses of automated trading systems. Rights infringement. The pervasive application of AI in mass surveillance and extensive data collection practices is increasingly encroaching upon basic human rights. Privacy breaches have become alarmingly routine as AI systems gather and analyze personal data on an unprecedented scale. This constant monitoring undermines the right to privacy, a cornerstone of individual freedom. Furthermore, these surveillance mechanisms exert a chilling effect on freedom of expression. Individuals, aware of the omnipresent AI-driven surveillance, may self-censor or refrain from expressing dissenting opinions, leading to a stifling of public discourse and democratic engagement. The situation is compounded by AI systems that enable discriminatory profiling and unwarranted scrutiny of individuals. Such practices, often lacking transparency and accountability, lead to systemic unjust treatment and exacerbate existing societal inequalities at scale.

#### AGI blows up capitalism – it causes massive wealth inequality, zero labor value, and stagnation

Stiefenhofer 25 [Pascal Stiefenhofer, researcher at the Department of Economics at Newcastle University, 02-10-2025, “Artificial General Intelligence and the End of Human Employment: The Need to Renegotiate the Social Contract”, ArXiv, https://arxiv.org/abs/2502.07050]/Kankee

Abstract The emergence of Artificial General Intelligence (AGI) labor, including AI agents and autonomous systems operating at near-zero marginal cost, reduces the marginal productivity of human labor, ultimately pushing wages toward zero. As AGI labor and capital replace human workers, economic power shifts to capital owners, resulting in extreme wealth concentration, rising inequality, and reduced social mobility. The collapse of human wages causes aggregate demand to deteriorate, creating a paradox where firms produce more using AGI, yet fewer consumers can afford to buy goods. To prevent economic and social instability, new economic structures must emerge, such as Universal Basic Income (UBI), which redistributes AGI-generated wealth, public or cooperative AGI ownership, ensuring broader access to AI-driven profits, and progressive AGI capital taxation, which mitigates inequality and sustains aggregate demand. Addressing these challenges in form of renegotiation the Social Contract1 1 The Social Contract, originally titled On the Social Contract; or, Principles of Political Right (French: Du contrat social; ou, Principes du droit politique), is a 1762 work by Swiss philosopher Jean-Jacques Rousseau. In this seminal text, Rousseau examines the nature of political legitimacy and proposes a model in which authority is derived from the collective will of the people, ensuring both governance and individual freedom. The book addresses the tensions of commercial society, a theme Rousseau had previously explored in Discourse on Inequality (1755), and offers a vision for a just and equitable political order based on social cooperation and mutual obligation. is crucial to maintaining economic stability in a post-labor economy. 1Introduction The rapid emergence of Artificial General Intelligence (AGI) marks a paradigm shift in production, labor dynamics, and economic power[7]. Unlike past technological advancements, which primarily enhanced human productivity, AGI possesses the capability to fully replace both cognitive and physical labor [8]. This unprecedented shift threatens to render human employment obsolete across numerous industries. Operating at near-zero marginal cost and continuously improving through self-learning, AGI offers firms an overwhelmingly efficient alternative to human workers. As a result, labor demand is poised to collapse, triggering a downward spiral in wages and fundamentally disrupting the historical equilibrium between labor and capital. In this new economic order, ownership of AGI assets—rather than human labor—becomes the primary determinant of wealth and power. The implications of this transformation are profound. Should human wages approach zero, traditional mechanisms of wealth distribution and economic participation become unsustainable. The classical capitalist model, predicated on wage-based consumption, faces a paradox: firms achieve unprecedented productivity while consumer purchasing power erodes due to mass unemployment. This dynamic threatens to destabilize markets, deepen economic inequality, and create a stark divide between AGI capital owners and those excluded from economic participation. The collapse of wage-based employment presents an urgent question: how should society adapt to an economy where human labor is no longer the primary means of income distribution? This paper explores the economic ramifications of AGI-driven automation and the policy interventions necessary to prevent systemic collapse. To ensure that AGI’s productivity gains benefit society as a whole rather than an elite minority, the existing social contract must be reimagined[9]. Without proactive policy responses, unchecked AGI-driven automation risks economic stagnation, extreme wealth polarization, and widespread social disruption. A critical lens for analyzing these transformations is the Cobb-Douglas production function, a foundational model in production economics that captures the relationship between labor, capital, and output [4][3][5]. Characterized by its elasticities—represented as exponents—the function quantifies how changes in input levels affect output. While widely applied in both economic theory and empirical modeling [5][6][11, 12, 14, 13], the Cobb-Douglas framework has yet to be extended to account for AGI’s impact on labor demand. As AGI becomes deeply integrated into production processes, its ability to dynamically adjust these elasticities in real time could significantly enhance total factor productivity TFP, improving overall efficiency and addressing resource misallocation [10]. This paper develops a novel production model that incorporates AGI-driven capital and labor. Section 2 introduces the foundational model integrating AGI capital, while Section 3 extends the framework to include AGI labor. Section 4 examines the combined effects of AGI labor and capital on production dynamics. Section 5 analysis the powershift towards AGI capital owners. Finally, the paper concludes with a discussion on broader economic implications and the policy measures required to navigate this unprecedented shift in the labor economy. 2Model I: Production Model with Human Labor, Capital and AGI as a form of Capital With the advancement of , the fundamental nature of production is shifting. Traditionally, economic models distinguish between capital ( K ) and labor ( L )[3]. However, as AGI becomes capable of performing cognitive and physical tasks at a superhuman level, it functions not as labor but as a productive asset owned and deployed by firms. Thus, AGI can be considered capital rather than labor, leading to human redundancy in the production process. In economic theory, capital is defined as any durable input used in the production of goods and services. AGI meets the criteria for capital due to the following properties. Definition: AGI can be defined as a new form of capital rather than labor, characterized by its ownership, control, and economic function as a deployable asset rather than a wage-earning entity. Like traditional machinery, AGI-based autonomous systems undergo depreciation and require continuous investment, updates, and maintenance to sustain functionality. It enhances productivity by optimizing decision-making, research, and production processes, thereby improving capital efficiency. Unlike human labor, AGI operates without wages, benefits, or rest, further aligning with the economic definition of capital. Given these attributes, AGI should be classified within economic models as part of the capital stock rather than as a substitute for human labor.

#### Advanced AI causes massive inequality and job losses

Haggstrom 16 [Olle Haggstrom, professor of mathematical statistics at Chalmers University of Technology, 2016, “Here be Dragons Science, Technology and the Future of Humanity,” Oxford University Press, https://academic.oup.com/book/26977]/Kankee

Frey and Osborne (2013) offer a systematic study of 702 different occupations, according to classifications by the US Bureau of Labor. A variety of standardized characteristics of the occupations serve as input to a statistical model producing, for each occupation, a number representing how susceptible the occupation is to robosourcing.227 While the model is crude and leaves out many potentially important factors (it doesn’t even explicitly involve any concrete mechanism or time dynamics), it may nevertheless give some hints as to which sectors are likely to be affected more and sooner than others. The authors offer the following summary statement: Our model predicts that most workers in transportation and logistics occupations, together with the bulk of office and administrative support workers, and labour in production occupations, are at risk. These findings are consistent with recent technological developments documented in the literature. More surprisingly, we find that a substantial share of employment in service occupations, where most US job growth has occurred over the past decades . . . are highly susceptible to computerisation. Additional support for this finding is provided by the recent growth in the market for service robots . . . and the gradual diminishment of the comparative advantage of human labour in tasks involving mobility and dexterity. Machines replacing human labor is of course not a new phenomenon. It has been especially prominent since the industrial revolution, but we have always found new tasks for human workers at about the same rate as machines have taken over the old ones (allowing for some variation in the booms and recessions of the economy). Today, however, when things are happening faster than ever, and when advances in artificial intelligence (AI) cause the automation not only of manual labor, but also of an increasing number of increasingly advanced intellectual tasks,228 there is good reason to revisit John Maynard Keynes’ prediction of widespread unemployment “due to our discovery of means of economising the use of labor outrunning the pace at which we can find new uses for labor.”229 Might the failure of unemployment levels in the United States and the European Union in the last few years to find their way back down to the levels from before the 2008 financial crisis be an early sign of Keynes’ prediction of rising levels of technological unemployment finally coming true? Working out the effects of robosourcing on unemployment is (like the case of outsourcing) complicated. In particular, the idea of analyzing each sector separately is too naive, because technological progress in one sector can cause unemployment in that sector but at the same time alter the supply-and-demand balance in other sectors, so that workers can find jobs there instead.230 The whole notion of technological unemployment has long been under a cloud among economists, and gone under the name “luddite fallacy,” but Brynjolfsson and McAfee (2014) give several reasons why, due to ways things are different now compared to, e.g., in the days of Keynes, we should expect climbing unemployment levels. One such reason is that an equilibrium analysis of the job market is not enough. There is a dynamic where workers in one sector are hit by technological unemployment, adapt, and find jobs elsewhere (in sectors where new business ideas are invented), but it may simply turn out that workers’ ability to adapt eventually fails when technological development goes too fast. There is a more philosophical issue that needs to be mentioned. We are used to talking about unemployment as a bad thing, but there is something a bit strange about doing so. On a fundamental level, our liberation from the hardship of labor can be seen a good thing, letting us focus instead on art, culture, sports, love or whatever we wish to fill our lives with. This is a very utopian vision, however, almost at the level of the transhumanist visions of Bostrom (2008a) quoted in Section 3.1. In practice, there are (at least) two things that should worry us. First (in the short to medium term), can the gradual transition to such a utopia be accomplished without negative social consequences of monstrous proportions? Second (in the long term), since society is and always has been organized very much around labor, the question arises whether we can organize society and our lives in such a way that the lack of work does not fill us with a sense of meaninglessness?231 Leaving that second question for the reader to ponder, let me mention some reasons why the first can be a real cause for concern. There are several mechanisms that suggest that our rapid technological development will (in the absence of government interference) tend to increase economic inequality. A basic one is that when machines take over jobs from humans, the income generated for the work goes from being income from labor to being income from capital (the money goes to the owner of the machine). The latter tends to be more unevenly distributed than the former. A second mechanism is that the typical impact of robosourcing is not the total elimination of jobs in a given category. Rather, what happens is that a large percentage of the jobs are lost, while the productivity of those remaining is greatly increased, often resulting in higher wages for those fortunate enough to hang on.232 A third mechanism is that globalization and digitization create a winner-takes-all economy: when the cost of making copies and distributing worldwide becomes negligible, competition hardens. In an old-style job like selling hot dogs, it may still suffice to provide the best service in your neighborhood, but in a modern occupation like constructing iPhone apps, you will probably be squashed if there is someone on the other side of the planet making a product that does the same thing as yours but a bit better.233 4.5 Intelligence explosion

#### Superior AGI cyberwarfare capabilities destroy global infrastructure and economies

Bikkasani 24 [Dileesh Chandra Bikkasani, researcher at the University of Bridgeport, 07-19-2024, “Navigating Artificial General Intelligence (AGI): Societal Implications, Ethical Considerations, and Governance Strategies, https://www.preprints.org/manuscript/202407.1573/v1/download]/Kankee

6.6. Cybersecurity Risks AGI presents profound cybersecurity implications. It could be leveraged to develop sophisticated cybersecurity attacks that are highly adaptive, which makes them harder to detect than current attacks. Given their advanced capabilities, these systems could rapidly scan for vulnerabilities and adapt to security measures, simultaneously launching coordinated attacks across multiple systems. AGI-powered attacks pose a significant threat to critical infrastructure like transportation networks, power grids, and communication systems and potentially cause widespread disruptions and failures (Raimondo et al. 2022). Robust security measures designed for AGI systems must be in place to defend against these advanced threats. Cybersecurity risks from AGI on critical infrastructure are particularly severe regarding public safety, national security, and economic stability. The potential compromise of critical infrastructure due to a cyber-attack could be devastating. This was evident in the case of “notPetya” where a cyber- attack on Maersk resulted in malicious software disrupting a fifth of the global shipping capacity, causing $10 billion in damage (Greenberg 2018). In another instance, the ransomware software “wannacry,” a self-propagating malware that encrypts victims’ data, causing a worldwide catastrophe impacting hospitals and other critical institutions (Chen and Bridges 2017). The estimated cost of cybercrime worldwide would skyrocket with the leverage of technologies like generative AI and AGI. 9 7. Legal and Policy Implications

#### AGI cyberattacks cause extinction – irreversible system failures, distrust, and economic collapse

Kasirzadeh 24 [Atoosa Kasirzadeh, tenure track Assistant Professor with joint affiliations in the Philosophy and Software & Societal Systems departments at Carnegie Mellon University with a Ph.D. in Philosophy of Science and Technology from the University of Toronto and a Ph.D. in Mathematics (Operations Research) from the École Polytechnique de Montréal, 2-6-2024, "Two Types of AI Existential Risk: Decisive and Accumulative", Arxiv, https://arxiv.org/html/2401.07836v2]/Kankee

6 The perfect storm MISTER Consider the highly interconnected world of 2035 where the pervasive integration of AI and IoT technologies has transformed almost every aspect of daily life.16 Cities embody a higher level of automation as compared to today, various functions in domestic and industrial sectors are automated, and even the most mundane devices like mirrors and refrigerators have become part of a vast, data-exchanging network. Frontier AI systems have become the backbone of this connected world. However, beneath the surface of this technological connectivity, vulnerabilities and risks have been brewing. AI-driven manipulation. The capability of AI to create convincing deepfakes reaches a critical height, where distinguishing fact from deepfake-generated reality becomes nearly impossible.17 This capability opens the door to the creation of hyper-personalized, AI-generated propaganda and similar persuasive narratives, which can be strategically leveraged for social engineering purposes — manipulating public opinion, influencing electoral outcomes, inciting civil unrest, diminishing oppositional viewpoints, and exploiting existing social and epistemic divides. Proliferation of Insecurity threats. In the domestic sphere, everyday personal IoT devices like smart mirrors and refrigerators transcend their roles as mere conveniences. They now represent significant points of vulnerability. With frontier AI capabilities, cybercriminals can penetrate these devices in a more expansive way, leading to widespread identity theft and ushering in a new era of digital espionage. What were initially perceived as isolated incidents of breaches gradually coalesce into a discernible pattern, signifying a more amplified erosion of digital security. The expansion of IoT devices also paves the way for the creation of extensive, interconnected botnets. These AI-powered networks, once relatively benign, show agentic abilities and become capable of launching unprecedented Distributed Denial of Service attacks, against critical infrastructures, including national power grids and communication networks. They represent a significant leap in cybersecurity risks. Each attack, incrementally more sophisticated than the last, signifies a disturbing escalation from individual cybersecurity concerns to widespread threats against national security. Moreover, as AI technologies becomes more widely available, they facilitate the emergence of new forms of bioterrorism. Private research labs with minimal expertise in synthetic biology and chemistry are now using AI to develop more infectious and deadly pathogens. The dual-use nature of AI in biotechnology — its potential for both beneficial and harmful applications — initially envisioned for medical breakthroughs, is maliciously repurposed to engineer biological weapons.18 Global mass surveillance and erosion of trust. By 2035, the implications of AI-induced mass surveillance have profoundly transformed socio-political structures worldwide. The earlier instances, such as China’s social credit system and the controversial NSA’s mass data collection, were just the precursors to a now-global trend. Governments across various political spectra have embraced advanced AI for mass surveillance, frequently under the guise of national security. This landscape is marked by an array of pervasive surveillance systems, some overt and others clandestinely operated under national security pretexts. In this future context, the reach of government surveillance have gradually eroded the sanctity of privacy even in putative democracies. The constant monitoring and evaluation of citizens by AI systems instill a culture of self-censorship among citizens, diluting the diversity of thought and social trust crucial for healthy democracy. In several countries, this leads to a radical erosion of trust in media as well as public and private institutions. Economic destabilization. In the financial market infrastructure, AI algorithms increasingly play a substantial role in analyzing and simulating market trends. This progression, however, introduces new challenges to economic stability. One emerging concern is the use of AI in creating sophisticated, but deceptive financial products. These products, while not outright phantom instruments, are complex derivatives or investment vehicles. It becomes increasingly difficult to distinguish these products from more traditional, stable investments, leading to potential market disruptions. The impact of these AI-enhanced financial products is particularly significant in already vulnerable economies facing systemic issues and cyber threats. The introduction of these complex investment tools further undermines trust in financial systems, intensifying the risk of economic downturns. Additionally, the potential for AI to unintentionally trigger market instability becomes a realistic concern. Automated trading algorithms, responding rapidly to market signals, now can inunprecedented speed amplify market volatility at unprecedented speeds. This phenomenon is exemplified in incidents like flash crashes — i.e., sudden and drastic market fluctuations driven by the high-speed responses of automated trading systems. Rights infringement. The pervasive application of AI in mass surveillance and extensive data collection practices is increasingly encroaching upon basic human rights. Privacy breaches have become alarmingly routine as AI systems gather and analyze personal data on an unprecedented scale. This constant monitoring undermines the right to privacy, a cornerstone of individual freedom. Furthermore, these surveillance mechanisms exert a chilling effect on freedom of expression. Individuals, aware of the omnipresent AI-driven surveillance, may self-censor or refrain from expressing dissenting opinions, leading to a stifling of public discourse and democratic engagement. The situation is compounded by AI systems that enable discriminatory profiling and unwarranted scrutiny of individuals. Such practices, often lacking transparency and accountability, lead to systemic unjust treatment and exacerbate existing societal inequalities at scale. In the perfect storm MISTER Scenario, a series of interconnected AI-induced risks coalesce into a catastrophic sequence of events, each exacerbating the next, leading to a potential existential crisis for humanity. The AI x-catastrophe unfolds with a devastating AI-driven cyberattack simultaneously targeting critical power grids across three continents. This orchestrated attack is the tipping point, a culmination of the escalating cybersecurity threats. The resultant continent-wide blackouts cause immediate and widespread chaos, disrupting essential services and plunging billions into darkness. The blackouts trigger a domino effect, causing major economic crashes. Financial markets, already destabilized by AI-induced manipulations, collapse under the strain. The economic fallout rapidly fuels societal unrest, with widespread protests and riots in response to the failing systems. Amidst this chaos and darkness, the seeds of distrust sown by AI-manipulated media, deepfakes, and targeted disinformation campaigns, which had been proliferating prior to the blackouts, begin to bear fruit. These divisive narratives, deeply entrenched in public consciousness, exacerbate social divides and impede efforts to restore stability and order. The blackout acts as a catalyst, propelling these latent tensions into active, widespread civil unrest. Simultaneously, the crisis exposes and amplifies previously minor inefficiencies and errors in AI systems, which become more pronounced due to the volatile market dynamics, regulatory upheaval, and ongoing algorithmic adjustments. These AI system failures extend their impact across various critical infrastructures, including healthcare and communication networks, further amplifying the societal disruption. The causal impact of AI inefficiencies limited to each subsystem, each seemingly non-existential in isolation, begins to accumulate dynamically and gives rise to compounded systemic impact, leading to disastrous global impacts. The convergence of these catastrophic events — multiple cyberattacks, manipulation, systemic eroded trust, economic destabilization, and rights infringements — leads to a state of global dysfunction. The capacity for a coordinated global response becomes critically undermined, as nations grapple with internal crises and widespread infrastructural breakdowns. Regional conflicts escalate into larger wars. Nations or non-state actors, driven by desperation or opportunism, engage in aggressive military actions, potentially leveraging AI technologies in warfare without legal constraints. In this scenario, the x-catastrophe emerges from the synergistic failure of systems critical to the functioning and survival of human civilization. The simultaneous and compounded nature of these crises creates a perfect storm situation where not only is recovery extremely challenging, but the potential for irreversible collapse is a stark reality. Before discussing the ramifications of the accumulative AI x-risk hypothesis, let’s first consider and address two potential objections. 7 Objections and replies

#### AGI implodes the carbon budget – computation power demands

Bikkasani 24 [Dileesh Chandra Bikkasani, researcher at the University of Bridgeport, 07-19-2024, “Navigating Artificial General Intelligence (AGI): Societal Implications, Ethical Considerations, and Governance Strategies, https://www.preprints.org/manuscript/202407.1573/v1/download]/Kankee

4. Implications for Energy and Climate 5 The development of AGI faces significant challenges in terms of energy consumption and sustainability. The environmental and ecological impacts are often overlooked when it comes to the advancements in this space. The computational power required to sustain AI models doubles every 100 days (Zhu et al. 2023). Increasing the capacity of a model by tenfold can result in a 10,000-fold rise in power demand. As AI systems become more advanced, their computational demands for training and running the system also increase, refer to Figure 2. Initiatives such as the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM-Global) by the United Nations Industrial Development Organization (UNIDO) highlight the importance of aligning AI advancements with global sustainability goals, particularly in mitigating the environmental impacts associated with AI and AGI technologies (UNIDO 2023). The two phases of energy consumption for AI systems are training and inference, with training consuming around 20% and inference consuming 80% of the resources. The energy demand for Natural Language Processing (NLP) tasks is exceptionally high, as the models have to be trained on vast datasets. Researchers estimated that training GPT-3 would have consumed around 1300MWh (Patterson et al. 2021). In comparison, GPT-4 is estimated to have consumed 51,772- 62,318 MWh of energy, which is roughly equivalent to the monthly output of a nuclear power plant. Additionally, the AI models have a significant carbon footprint, with some transformer neural networks emitting over 626,000 lbs. of CO2 (Kurshan 2023). Techniques such as power-capping the GPUs would reduce energy usage by 15% and only a marginal increase in computational time (McDonald et al. 2022). Another promising avenue is the development of AI-specific energy-efficient hardware designed for workloads. Companies like Nvidia, Google, and Intel are investing in AI chips and tensor processing units (TPUs) that deliver high performance while consuming less power than CPUs and GPUs (Sze et al. 2020). Distributed and federated learning are also being considered to distribute the computing load across multiple devices or edge nodes, reducing the energy demands on centralized data centers (Konečný et al. 2016). 6 If current language models require such immense amounts of energy and computational power for training and inference, developing AGI would necessitate far more resources, leading to even more significant environmental impacts on society. The pursuit of AGI would demand orders of magnitude more computational resources than current narrow AI models. As the AGI systems increase in scale and complexity, the energy requirement and carbon footprint would escalate exponentially, potentially straining existing energy infrastructure and exacerbating climate change concerns. Addressing sustainability, energy efficiency, and their challenges will be crucial to mitigate its societal and ecological consequences (Ammanath 2024). Shifting towards renewable energy sources and energy-efficient computing infrastructure is essential to minimize the associated environmental impact. Leveraging renewable resources like solar, wind, and hydroelectric power can significantly reduce the strain on current infrastructure and the carbon footprint, aligning with global efforts to combat climate change. 5. Ethical Implications

#### Unrestrained advanced AI development collapses planetary life support – carbon, water, and pollutants

Bhardwaj et al. 25 [Eshta Bhardwaj, researcher at the University of Toronto, Rohan Alexander, researcher at the University of Toronto, and Christoph Becker, researcher at the University of Toronto, 01-2025, “Limits to AI Growth: The Ecological and Social Consequences of Scaling,” Arxiv, https://arxiv.org/pdf/2501.17980]/Kankee

Academic capture indirectly influences research agendas [187, 199] and generally takes the form of funding institutions, sponsoring academic conferences or events, inviting certain academics as research experts on panels or boards, or providing compute infrastructure and resources [187]. This is done by these companies to incentivize and influence certain types and topics of research as well as for self-promotion. Beyond capture, it is seen that AI companies demonstrate ethicality in a performative manner to appease the larger public. Ethics washing (“the practice of feigning ethical consideration to improve how a person or organization is perceived” [1]) is one mechanism used to counter and sidestep increased organized resistance while influencing regulatory functions. Ethics shopping is “the malpractice of choosing, adapting, or revising ethical principles... from a variety of available offers, in order to retrofit some pre-existing behaviours...” [50, p. 186]. This prevents comparisons between companies and makes it more difficult to parse as the language is mixed and inconsistent. Ethics dumping occurs when companies export “...unethical research practices to countries where there are weaker legal and ethical frameworks and enforcing mechanisms.” [50, p. 189]. This is seen particularly with AI data work, such as in the case of business process outsourcing (BPO) companies in Africa who employ data workers that have to parse through disturbing and gruesome content while working under contracts that incentivize productivity targets like speed and accuracy [138, 156]. Many companies offset emissions by purchasing carbon credits that invest in carbon removal and emissions reductions projects, e.g. [43, 47, 58, 98, 99, 123, 169]. While these appear to mitigate carbon emissions, ample evidence shows offset credits to be misleading attempts at greenwashing [10, 30, 62, 83, 131, 140, 144]. For example, an investigation into Verra revealed 90% of the rainforest offset credits to be questionable, and 94.9 million carbon credits were claimed for only 5.5 million real emissions reductions [62]. A more comprehensive study of over 2000 projects shows that only 16% of carbon credits lead to real emissions reductions [140]. Despite all these captures, then, the ecological and social costs of AI growth have become hard to ignore as limits to growth become increasingly acute. 5.3 Limits to Growth, Overshoot, and Collapse Returning to the theme of Limits to Growth, consider Fig. 8. The only difference to Fig. 1 is an added link between the state of the system and the carrying capacity. This link is based on the recognition that system activity itself can affect the capacity of the containing ecosystem to support continued activities [120], such as ecosystem viability and planetary life support systems [64, 147, 149]. In the anthropocene, computing has reached activity levels with measurable impact on a planetary scale [34]. The crucial difference is that unlike a limited growth scenario in which the activity level of a system stagnates, like a saturated market, the erosion of carrying capacity undermines the continued stability of the entire ecosystem and leads either to system collapse or an unstable oscillation trajectory. Overshoot and collapse patterns can apply in at least two ways. First, model collapse already has been proven with respect to deteriorating data quality, as discussed earlier. In this case, resource adequacy would refer to the availability of high-quality training data for LLMs. The proliferation of generated text online erodes that availability as already illustrated by the closure of Wordfreq noticed above [96]. The collapse of model quality by continued erosion of data quality has been demonstrated empirically [72, 166]. Second, if the generative AI boom continues to drive excessive investments in new hyperscale data centers, their material impact will impact ecosystems on a global scale, significantly affecting planetary boundaries including atmospheric CO2, air quality, freshwater cycles, and pollution [150, 171]. Alternative paths can be opened by refocusing. A moderate refocusing could involve moving from resource-intensive progress on model performance to resource-efficiency progress as proposed recently [51, 182]. While this is presented as frugal AI, it should be noted that a more radical refocusing on frugality would shift the focus altogether from efficiency to sufficiency by placing primary emphasis on asking “what’s good enough?”. Nevertheless, frugal AI is a promising avenue for course correction that would much more effectively address limits to growth since it focuses on the primary drivers of accelerating scaling. In other words, it is a high leverage point for system change [119] to open spaces for genuinely innovative development. 6 Conclusion Above, we have mapped out the accelerating growth of AI along the lines of technical, economic, ecological, and social dimensions and their interactions. We have illustrated how the mathematical laws of scaling translate into technical development with economic implications of scale subject to capitalist and market dynamics. The resulting economies of scale cause ecological destruction, which is ultimately an ethical concern too. We explored the dynamics of how these perspectives interact through causal modelling. Drawing on system dynamics archetypes, we argue that the AI industry’s responses to barriers typically attempt to overcome apparent limits in one perspective but fail to account for resulting damages in other perspectives. These damages cause social and ecological harms that are externalized by the AI industry but demonstrate a vital need to realign our priorities around scaling. Lastly, we emphasize the need for refocusing on sufficient AI practices to avoid an overshoot and collapse trajectory. Growth always ends – either by design, or by disaster. Which path we choose has not been decided yet. In Fall 2024, the United Nations Pact for the Future enshrined a commitment to develop measures of progress beyond GDP growth in future agendas and sustainable development policy [177]. As the world comes to finally recognize that pursuing relentless growth is the path to assured disasters, it is time that the AI industry catches on and changes course.

#### High-end AI models are an ecological disaster

Bhardwaj et al. 25 [Eshta Bhardwaj, researcher at the University of Toronto, Rohan Alexander, researcher at the University of Toronto, and Christoph Becker, researcher at the University of Toronto, 01-2025, “Limits to AI Growth: The Ecological and Social Consequences of Scaling,” Arxiv, https://arxiv.org/pdf/2501.17980]/Kankee

3.3 Ecological While the list of concerns is broad and includes waste, air pollution, minerals, and the impact of mining, below we focus on energy, water, and CO2 emissions to illustrate the ecological perspective. AI requires energy and materials to manufacture, power, run, and cool hardware. Dedicated AI data centers are increasingly specialized for training (data processing and ML model training) and inference (deployed models used in applications) [158]. While inference is tiny when compared to training, it occurs much more frequently [84, 155] and its energy demand varies based on task complexity (e.g., image generation versus text classification [113]). In the case of the BLOOMz-7B model, approximately 590 million inferences would be needed for the energy cost of inference to match that of training [113]. But for popular models like GPT with 100 million monthly users [115], inference is so frequent that it heavily outweighs the energy impact of training [113]. About 50% of energy consumption of data centers is used to power the hardware and another 40% to cool it [114]. For example, the energy demand of the LLaMa 3.1 405B model (released in July 2024) for powering and cooling peaks at 1.7kW per GPU [4]. At 16,000 GPUs, this amounted to roughly 27.2MW [162]. Newer chip generations have also steadily increased power consumption (Fig. 3d shows the tripling of maximum power consumption enabled by NVidia chips within four years). While clean water is needed to cool data centers and to generate electricity (through water-intensive thermoelectric plants) [114, 165], chip manufacturing requires “ultrapure” water [73, 105] and one facility “can use 37 million litres of ultrapure water per day” [93]. Google’s reported data center water consumption rose from 16 billion litres in 2021 to 28 billion in 2023 [60, 179] and it’s been found that larger models need more water (see Fig. 3e). CO2 is emitted at all stages of the AI lifecycle. For the BLOOM language model, “of the total 50 tonnes of CO2eq emissions emitted during model training, only half was due to the energy consumption of the GPUs used for training BLOOM (‘dynamic consumption’), with a further 29% stemming from the idle consumption of the data center (i.e., the energy used for heating/cooling, networking, storage, etc.), and a final 22% ...from the GPU manufacturing process.” [114, p. 6]. Aggregate emissions are increasing over time as models become larger (see Fig. 3f), require more data centers, and need chips with greater processing power. Google’s emissions increased nearly 50% from 2019 to 2023, Microsoft’s emissions increased 29% from 2020 to 2023, and Meta’s increased 66% from 2021 to 2023 [57]. Efforts towards compute efficiency also do not translate to energy efficiency and hence do not result in savings in carbon emissions [195]. Barrier 4: Shortage in energy supply. Building more data centers and developing new, advanced GPUs is energy-, material- and water-intensive and generates carbon emissions and other pollutants. Data center capacity (in megawatts) has seen nearly a 2000% increase in the U.S. alone from 2019 to 2024 [45]. Of particular concern to the AI industry is the shortage in the power supply that places constraints on further expansion due to pre-existing contractual commitments for energy and time requirements for expanding or acquiring new sources of it [162]. Barrier 5: Accountability for environmental impact. Companies have claimed sustainability goals such as to be carbon neutral, but AI advancements require rising electricity and water consumption. As for growing emissions [2, 9, 127, 191], companies have made future promises only to renege, postpone, or walk back on them, e.g., [71, 143]. While promises of sustainable practices continue, so does the development of increasingly complex models that require more energy, more water, and cause more emissions. These operations reveal business priorities. At the moment, real transparency and accountability are not in place, such as in the case of data centers that are not legally required to report their water usage figures while companies continue to lobby to keep this information a trade secret [53]. 3.4 Social AI’s Scale of Impact. The widespread adoption of AI poses old questions in new forms as well as new ethical challenges [31, 54, 56]. Its scale of impact is larger than ever before in the history of computing ethics due to the opacity of the models, the number of people it can impact, its potential for damage, and the feedback loops that increase their scale of harm [136]. A major consequence of the prevalence of AI in our society is the standardization and lack of autonomy in deciding whether you want to be part of the data economy [121] as power centralizes in the hands of monopolies driving the “social quantification sector” [32, 121, 122]. Datafication enables surveillance capitalism in a “transformation of human life so that its elements can be a continual source of data” [121, p. 2],[202]. The ability to datafy remains largely with the big AI companies. “Inequality resides... in having or not having the power to decide what kind of data is being generated... by whom, for what purpose, and for whose benefit.” [49, p. 831]. This also enables the exploitation of labour. Millions of gig workers, often employed from India, Kenya, Philippines, or Mexico, are paid $1.46/hour to enable the development of AI models [76]. The poor labour conditions of the workers include monitoring through automated tools [193]. The lack of equity in the actors that can afford and access data, and the monopoly they exercise signals that “...capitalism is premised upon the preservation of unequal power: of the enforcement of the racial, gendered, and other social hierarchies which enable the extraction of labor, and therefore value, from the many for the profit of the few.” [94, p. 102],[32]. Thus the computing industry offloads its ethical debts to distant stakeholders [19]. Environmental Justice. It has long been recognized that ecological and social impacts are closely linked, and environmental sustainability is always an issue of equity and social justice [8]. A few examples follow. The globally dominant chip manufacturer TSMC is estimated to consume as much electricity as New Zealand by 2030 [75]. The consequences were experienced by over 5 million residents impacted by blackouts caused by a grid operating much closer to capacity than it should [75]. Data centers impact communities globally [33, 80], including by distorting the electric grid and causing billions of dollars in damages to appliances [104] and by air pollution, which alone is estimated to cause 1300 deaths annually in the U.S. by 2030 [70]. In Uruguay, Google plans to construct a data center predicted to produce 25,000 tonnes of CO2 while releasing 86 tonnes of hazardous waste with no plans in place of how the waste matter will be disposed [108]. This construction is estimated to use 7.6 million litres of water at a time when Uruguay is experiencing a drought, water that could be used by 55,000 people [107]. Barrier 6: Organized resistance. In response to these environmental injustices, the erosion of worker autonomy, and injustices resulting from data colonialism [32], organized public resistance has increased [117], including residents mobilizing against data centers in many locations [42, 134], calls to defund Big Tech [16], environmental activism by Amazon workers [59], campaigns against military projects by Google employees [174], legal action by content moderators [25], and calls to democratize decision-making about how AI is developed and deployed [82]. 3.5 Dynamics of AI Scaling

#### Huge AI power demands revive fossil fuels

Riordan 25 [Michael Riordan, scientist with a doctorate from the Massachusetts Institute of Technology, 01-31-2025, "AI’s Energy Demands Threaten a Nuclear Waste Nightmare", Scientific American, https://www.scientificamerican.com/article/ais-energy-demands-threaten-a-nuclear-waste-nightmare/]/Kankee

Long in decline, the U.S. nuclear industry is hoping for resurrection at two sites of its greatest failures: Three Mile Island in Pennsylvania and the Hanford Site in Washington state. Nuclear power, the industry claims, will help satisfy the surging power demands from data centers and the growing AI economy. But such a wrong turn ignores the long-unresolved problems of radioactive nuclear wastes that AI cannot wish away. In September Constellation Energy announced plans to restart a shuttered reactor at Three Mile Island, prodded by Microsoft, which will need many gigawatts of power to perform extensive AI calculations in its expanding fleet of data centers. Amazon followed suit and announced in November that it will invest $334 million to develop small modular nuclear reactors (SMRs) at Hanford, site of the world’s first plutonium-production facility. Google and Meta are also hoping to bring nuclear power back. In October 2024 Google announced it eventually plans to purchase 500 megawatts of electricity from Kairos Power, which is developing a novel SMR in Oak Ridge, Tenn., on the site of the national lab that long refined uranium for the nuclear industry. And Facebook parent Meta is seeking bids for nuclear power plants for its data centers. These tech giants recognize that the next generation of microprocessors to be used for AI calculations at data centers will require oodles of electricity to power and cool them. A single Nvidia Blackwell chip, for example, can draw up to two kilowatts, more than what is needed for a typical house. Cram thousands of them in servers inside a data center, and they will need as much power as a small city. So-called hyperscale data centers require over 100 megawatts (100 MW)—a sizeable fraction of the output of a major power plant. And that power should be cheap, steady and reliable. An authoritative December 2024 report from the U.S. Department of Energy, written by energy experts at the Lawrence Berkeley National Laboratory, is especially illuminating. The growth in U.S. data-center energy usage over the next five years, they state, would correspond “to a total power demand for data centers between 74 and 132 [gigawatts].” That would represent some 7 to 12 percent of the U.S. electricity consumption forecast for 2028. Where on Earth is all this power going to come from? Given the challenges electric utilities face in supplying electricity to meet other growing needs, including electric vehicles, it’s small wonder that big tech has turned back to the atomic nucleus. But the power demands outlined in the DOE report would require building or resurrecting the equivalent of at least 40 Three Mile Island reactors over the next five years. That’s impossible. Several years ago Amazon, Google, Meta and Microsoft promised not to exacerbate atmospheric carbon-dioxide levels. But that laudable goal is becoming increasingly difficult to achieve, given their data centers' exploding electricity consumption. So they have instead begun touting a return to nuclear power to avert this thorny problem. That’s a huge mistake. Nuclear power is indeed a source of carbon-emission-free energy, but it is hardly a clean energy source, and it is definitely not renewable. All along the uranium supply chain—from mining to enrichment to the fabrication of fuel rods or pellets—opportunities abound for radioactive releases. In South Texas, for example, landowners worry about contamination of their groundwater by renewed uranium mining activities nearby. Since 1989, the DOE has spent hundreds of billions of taxpayer dollars cleaning up the original nuclear complex—including the gargantuan Oak Ridge factory that enriched much of the uranium used for commercial nuclear power. And despite decades of trying, the department has yet to fully clean up and dismantle the oozing, disintegrating tanks of highly radioactive wastes left over from plutonium processing at Hanford. The storage and containment of spent nuclear fuel is in fact the crucial unresolved challenge of the U.S. nuclear industry. Over 90,000 tons of these wastes are stored at 77 sites in 35 states—an amount increasing by over 2,000 tons a year. Small modular reactors, promoted by Microsoft co-founder Bill Gates and others, will only add to this growing burden. As former U.S. Nuclear Regulatory Commission (NRC) chair Allison Macfarlane and Rodney C. Ewing of Stanford University stated, “In some cases these new reactors may make it worse by creating more waste that’s more costly to manage, new kinds of complex waste, or just more waste, period.” Elsewhere, Macfarlane stressed the procedural and practical difficulties faced by novel nuclear reactor technologies in gaining NRC acceptance and achieving commercial success. Shortly after it had received NRC certification in 2023, for example, the much-touted NuScale SMR project was abandoned after anticipated construction costs more than doubled to $9.3 billion. Leaving aside the waste problem, a commercially successful SMR design is probably over a decade away. But the relentless AI gold rush, if left unchecked, will impose unattainable demands on projected power supplies well before that. Meanwhile, electricity rates will rise inexorably in light of the law of supply and demand. That looming energy crisis explains big tech’s efforts to slow shutdowns of fossil-fueled power plants and to resurrect shuttered reactors. Amazon, Google, Meta and Microsoft executives should instead take a deep breath and begin reevaluating their options. Do they really need to build and upgrade data centers at such a breakneck pace? Or is this devil-take-the-hindmost AI arms race just the result of bitter competition, prodded by recent advances in semiconductor technology? And what about the truly clean, renewable energy sources they once embraced—especially solar, wind and geothermal? Yes, the variability of solar and wind energy makes them a poor match to the steady power requirements of data centers. But energy storage has come a long way recently and has a promising future. And the recent startling success of the Chinese DeepSeek AI program demonstrates that software efficiency will play an important role in this effort. Given the dark clouds still lingering over nuclear power, especially its unresolved waste problem, these renewable alternatives deserve renewed consideration.

#### AGI takes hundreds of years’ worth of energy

Winfield 20 [Alan F.T. Winfield, Professor of Robot Ethics at UWE Bristol, Honorary Professor at the University of York, Associate Fellow in the Cambridge Centre for the Future of Intelligence, and chair of the advisory board of the Responsible Technology Institute with a BSc and PhD in electronic engineering, University of Oxford2020, “ Chapter 1 On the simulation (and energy costs) of human intelligence, the singularity and simulationism,” From Astrophysics to Unconventional Computation, https://link.springer.com/chapter/10.1007/978-3-030-15792-0\_16]/Kankee

1.2.2 The energy cost of evolving human-equivalent AI In the quest for human-equivalent AI there are, broadly speaking, three approaches open to us: design it, reverse-engineer it2 or evolve it. The third of these – artificial evolution – is attractive because it sidesteps the troublesome problem of having to understand how human intelligence works. It’s a black box approach: create the initial conditions then let the blind watchmaker of artificial evolution do the heavy lifting. This approach has some traction. For instance David Chalmers, in his philosophical analysis of the technological singularity [4], writes “if we produce an AI by artificial evolution, it is likely that soon after we will be able to improve the evolutionary algorithm and extend the evolutionary process, leading to AI+”. And since we can already produce very simple AI by artificial evolution, then all that’s needed is to “improve the evolutionary algorithm”. If only it were that straightforward. Some time ago I asked myself (and anyone else who would listen): ok, but even if we had the right algorithm, what would be the energy cost of artificially evolving human-equivalent AI? My hunch was that the energy cost would be colossal; so great perhaps as to rule out the evolutionary approach altogether. That thinking, and some research, resulted in a short paper in ALIFE 14 [16]. That paper explores the question: what is the energy cost of evolving complex artificial life? The paper takes an unconventional approach by first estimating lower and upper bounds on the energy cost of natural evolution and, in particular, the species Homo Sapiens Sapiens. A lower bound of ∼ 8000EJ3 is obtained by estimating and summing the calorific energy costs of populations of 7 concestor species stretching back from hominids to singlecelled organisms over about 3.5B years. The upper bound of ∼ 5.7 × 1012EJ is obtained by simply estimating the amount of solar energy available to the evolution of all living things. Even though these are very crude estimates they have value because we are forced us to think about the energy costs of co-evolution, and hence the energy costs of evolving complexity. Of course the processes and mechanisms of biological and artificial evolution are profoundly different (except for the meta-level equivalence of the Darwinian evolutionary operators: variation, selection and heredity), but there is an ineluctable truth: artificial evolution still has an energy cost. Virtual creatures, evolved in a virtual world, have a real energy cost. And we can estimate that energy cost. For a simple robot with an artificial neural network of comparable size to that of the nematode worm C. elegans (with 302 neurons), each simulated robot has a real energy cost of about 9J/hr, which interestingly is about 2000 times greater than the energy cost of a very small (1mg) organism, 0.004J/hr. It is clear that ‘larger’ artificial creatures, i.e. with more artificial neurons, must incur a greater computational energy cost. In general, if the energy cost of simulating and fitness testing a virtual creature is e, then the energy cost of evolving that creature will be E = gpe, where g is the number of generations required and p the population size. Energy cost e is clearly a function of the complexity of that virtual creature, but how might e scale with complexity? Kleiber’s law [5, p. 422] relates the mass of an organism to its energy consumption and, plotted on logarithmic axes, shows a remarkably consistent linear relationship from micro-organisms to the largest animals. Perhaps a similar relationship might exist between, say, neural complexity and energy cost e for virtual creatures: an artificial life equivalent of Kleiber’s law? Figure 1.1 imagines such a plot, of neural complexity against energy cost e. We cannot yet plot such a relationship since we have, to date, only one or two points at the very bottom of the artificial neural complexity scale. But, if we assume that a human-equivalent AI will require roughly comparable neural complexity to Homo Sapiens4 , with 85 × 109 neurons and 1014 − 1015 synapses (and noting that neural complexity must take account of the number of synapses, since neural connections incur a computational energy cost), then e for an artificial creature of this synaptic complexity could be 1010 − 1012 times greater than for something equivalent to C. elegans. But this scale factor is still likely to be too low because fitness testing of increasingly complex artificial creatures will take longer and incur greater energy cost. It seems likely that the gradient of our ALife version of Kleiber’s law will be greater than 1. I estimate that the computational energy cost of simulating and fitness testing something with an artificial neural and synaptic complexity equivalent to humans could be around 1014KJ, or 0.1EJ. But evolution requires many generations and many individuals per generation, and many co-evolving artificial species. Also taking account of the fact that many evolutionary runs will fail (to produce smart AI), the whole process would almost certainly need to be re-run from scratch many times over. If multiplying those population sizes, generations, species and re-runs gives us (very optimistically) a factor of 1,000,000 - then the total energy cost would be 100, 000EJ. In 2010 total human energy use was about 539EJ. So, artificially evolving human-equivalent AI would need the whole human energy generation output for about 200 years. 1.3 The Technological Singularity

#### AI development is environmental racism

Bender et al. 21 [Emily M. Bender, Endowed Professor in the Department of Linguistics and the faculty director of the CLMS program and the director of the Computational Linguistics Laboratory at the University of Washington with a PhD from the Linguistics Department at Stanford University, Seattle, WA, USA Timnit Gebru, computer scientist with a Bachelor of Science and Master of Science degrees in electrical engineering and a PhD in computer vision, Angelina McMillan-Major, researcher with a Ph.D. in Computational Linguistics from the University of Washington, and Shmargaret Shmitchell, computer scientist with a PhD from the University of Aberdeen (also known as Margaret Mitchell), 2021, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” FAccT, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922]/Kankee

While [129] benchmarks the training process in a research setting, many LMs are deployed in industrial or other settings where the cost of inference might greatly outweigh that of training in the long run. In this scenario, it may be more appropriate to deploy models with lower energy costs during inference even if their training costs are high. In addition to benchmarking tools, works estimating the cost increase associated with the introduction of LMs for particular applications, and how they compare to alternative NLP methods, will be important for understanding the trade-offs. When we perform risk/benefit analyses of language technology, we must keep in mind how the risks and benefits are distributed, because they do not accrue to the same people. On the one hand, it is well documented in the literature on environmental racism that the negative effects of climate change are reaching and impacting the world’s most marginalized communities first [1, 27].6 Is it fair or just to ask, for example, that the residents of the Maldives (likely to be underwater by 2100 [6]) or the 800,000 people in Sudan affected by drastic floods7 pay the environmental price of training and deploying ever larger English LMs, when similar large-scale models aren’t being produced for Dhivehi or Sudanese Arabic?8 And, while some language technology is genuinely designed to benefit marginalized communities [17, 101], most language technology is built to serve the needs of those who already have the most privilege in society. Consider, for example, who is likely to both have the financial resources to purchase a Google Home, Amazon Alexa or an Apple device with Siri installed and comfortably speak a variety of a language which they are prepared to handle. Furthermore, when large LMs encode and reinforce hegemonic biases (see §§4 and 6), the harms that follow are most likely to fall on marginalized populations who, even in rich nations, are most likely to experience environmental racism [10, 104]. These models are being developed at a time when unprecedented environmental changes are being witnessed around the world. From monsoons caused by changes in rainfall patterns due to climate change affecting more than 8 million people in India,9 to the worst fire season on record in Australia killing or displacing nearly three billion animals and at least 400 people,10 the effect of climate change continues to set new records every year. It is past time for researchers to prioritize energy efficiency and cost to reduce negative environmental impact and inequitable access to resources — both of which disproportionately affect people who are already in marginalized positions. 4 UNFATHOMABLE TRAINING DATA

#### AGI arms races destroy deterrence, causes misaligned AI, and leads to runaway cyberwarfare

Yampolskiy and Ramamoorthy 18 [Roman Yampolskiy, researcher at the department of Computer Engineering and Computer Science at the University of Louisville, and Anand Ramamoorthy, researcher at the Dept. of Anaesthesiology at UW-Madison, 2018, “Beyond Mad?: The Race for Artificial General Intelligence,” ITU Journal, https://www.itu.int/dms\_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P09-PDF-E.pdf]/Kankee

2. ARMS RACES AND AGI: BEYOND MAD? The success of The Manhattan Project and the deployment of nuclear weapons by the United States military in the Second World War led to a new kind of international conflict, a nuclear arms race, where powerful nations sought to acquire the same destructive capabilities as the U.S.A. This resulted in a world where an unstable peace is kept alive, informed to a significant extent, by the doctrine of mutually assured destruction (MAD) in addition to global non-proliferation efforts. A detailed discussion of the current understanding of MAD and the status of nuclear non-proliferation is beyond the scope of this paper. It suffices to note that examining the case of MAD in its original context provides insights that can be used to understand the role of disruptive technologies in international conflict (for case studies of such technologies see [9]). AGI, if and when it happens, may well be the final disruptive technological development engineered by humans. AGI represents a level of power that remains firmly in the realm of speculative fiction as on date. It stands to reason that if true AI were achievable, state actors would be invested in achieving this and with priority if possible. Such a quest for priority might have disastrous consequences due to corner-cutting when it comes to safety, and has been described as “racing to the precipice” [10]. An AI arms race is often spoken of in the context of the development of autonomous weapons systems which become increasingly sophisticated, changing the face of warfare. Were we to adopt Clausewitz’s observation that “war is the continuation of politics and with other means” [11] and examine international conflict, it becomes obvious that the role of AI would extend well beyond, and emerge well before, armed conflict. A nation equipped with a fully general AI, would stand to benefit in the negotiation of conflict and agendas, regardless of means. If said AI were both general AND endowed with the ability to act in the world (i.e., not merely an Oracle-in-a-box as some have proposed, see [12] for an analysis of AI confinement), then, all arguments pertaining to the existential risk posed by AI would apply. Having AI systems autonomously determine the deployment of weapons in armed conflict is one major route to potential catastrophe, but we would like to emphasize that matters are likely to become fraught even before this development. AGI development would push the global security strategy beyond what is currently in place. In the event of human control over the AGIs (which is a problem worth examining in its own right), MAD would not be sufficient to avert catastrophe. This would be because of the greater complexity associated with AGI and the capabilities such a system would present to human controllers, for instance, the AGI of a particularly belligerent state could calculate the optimal means to mortally destabilize the economy of a rival state (however powerful), or develop weaponized code disseminated globally to control, if not disrupt, vital systems such as power grids and communication networks. In other words, the cyberwarfare capabilities of an AGI-assisted nation-state would pose a serious threat to global stability and humanity. The current forms of narrow AI are capable of interfering with communication services. AI enabled surveillance across communication networks is likely to become the norm. AI tools with the potential to perturb or alter the content of communications are already in development (see https://lyrebird.ai/ for an interesting example in the context of mimicking human speech; see also: https://lyrebird.ai/ethics/). An AGI with access to the Internet and communication networks in general would be able to, depending on its objectives (or of those who deploy it), selectively impede communication across a certain network/region, or fabricate misinformation to probe human responses if it develops the objective to understand social impact of communication networks. Much as these scenarios remind us of science fiction, it is worth noting that we encounter reports of computational propaganda or technology-assisted disinformation with increasing regularity. On a more optimistic note, an AGI that is constrained to cooperate with humans could help envision more efficient use of resources for optimizing the communication networks we have available, or design altogether novel and better architectures. As we discuss below, the potential development of AGI is unlikely to be an exclusively state-funded affair. Current breakthrough AI systems all appear to be products designed by technology companies. Given the fact that technology giants such as Google, Facebook and others are beginning to open-source their machine-learning tools (e.g. TensorFlow), it is well within the realm of possibility that non-commercial, non-state actors including individuals would be able to develop applied artificial intelligence. 2.1. Actors in the AGI race

#### AGI enables world-ending cyberterror

Yampolskiy and Ramamoorthy 18 [Roman Yampolskiy, researcher at the department of Computer Engineering and Computer Science at the University of Louisville, and Anand Ramamoorthy, researcher at the Dept. of Anaesthesiology at UW-Madison, 2018, “Beyond Mad?: The Race for Artificial General Intelligence,” ITU Journal, https://www.itu.int/dms\_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P09-PDF-E.pdf]/Kankee

2.1.3. Rogue actors Unlike other dangerous technological developments in history, AI breakthroughs may not occur exclusively in academic, governmental or industrial centers of research. In principle, powerful AI systems could be developed by individuals or groups with no national or corporate agendas. Such homebrewed AI would be hard to deal with precisely due to a lack of oversight, monitoring or consensus on architectures and objectives. More worrisome is the prospect of such rogue actors developing AI without safety considerations or with malicious intent (for example, see the case of commercially available unmanned aerial vehicles being repurposed for guerrilla warfare by terror groups [14]). It could be argued that the resources required to develop a powerful and truly general artificial intelligence may not be available to rogue actors, but it is far from clear that this would be the case, and it may be unwise to presume that any such obstacles would be insurmountable. Cyberattacks originating from individuals/small groups are commonplace in our increasingly interconnected world, and it is conceivable that the development of an artificial general intelligence by rogue actors would be similar in terms of execution, but more harmful in terms of impact on human society and life. As a case in point, consider the recent largescale spread of the WannaCry ransomware, exploiting a vulnerability in the Windows operating system (particularly versions past). Investigations seem to suggest that the architects of the attack were not well organized and the attack not as nightmarish as it could have been. Yet, it precipitated a significant amount of chaos and affected networked computers worldwide. Now replace the ransomware with an AGI that is released into the wild by hackers motivated by political ideology, notoriety or curiosity [15]. Even if the AGI is not inherently dangerous, the consequences in such a scenario would be hard to predict or plan for and could be catastrophic. 3. AGI AND VALUE ALIGNMENT

#### AGI-capitalism causes extinction – ecology and war

Dyer-Witheford et al. 19 [Nick Dyer-Witheford, associate professor at the University of Western Ontario in the Faculty of Information and Media Studies, Atle Mikkola Kjøsen, assistant Professor in the Faculty of Information and Media Studies at the University of Western Ontario, and James Steinhof, Assistant Professor / Lecturer and Ad Astra Fellow in the School of Information and Communication Studies with a PhD in Media Studies from the University of Western Ontario, 2019, “Inhuman Power Artificial Intelligence and the Future of Capitalism,” Pluto Press, https://www.jstor.org/stable/j.ctvj4sxc6]/Kankee

AI’s Dirty Secret In the loudly proclaimed ethical and safety-conscious deliberations of leading AI-capitalists, attention is now given to AI as an ‘existential risk’ (Bostrom 2014: 4). Such risk arises largely because of the possibility of an AGI evolving into an ASI beyond human control. The issue is not malevolence (Skynet) but rather efficiency. Nick Bostrom’s (2014: 123–5) famous example is of an AI instructed to make paper clips that attains superintelligence and uses its ever-extrapolating powers to convert the entire universe into paper clips, obliterating humanity as collateral damage. The point is a serious one, even if currently (we hope) remote; not only could ASI’s ‘integral accidents’ – Paul Virilio’s (2000) term for malfunctions so intrinsic to a given techno-system they must be considered a feature, not a bug – be devastating, they might not be ‘accidents’ at all. However, Bostrom’s prediction – and other similar warnings, such as that of nanotechnologist Eric K. Drexler’s (1987) earlier ‘grey goo’ scenario, in which out-of-control self-replicating nano-robots consume all biomass on Earth while building more of themselves – can also be interpreted in a wider sense. This is to understand them not literally but metaphorically, or rather both literally and metaphorically, as simultaneously identifying a concrete hazard and providing a parable of runaway economic growth and universal commodification. The real ‘paper clip’ is profit, and the manifest danger of AI is not only that of an ‘integral accident’ but equally or more of its intended use as a means of intensifying and accelerating the production and circulation of goods that is destroying the environment, annihilating species, and, for humans, heating the planet to civilizational, perhaps existential, limits. This untrammelled economic growth is of course the very profit-maximizing process that would lie behind an unlimited order of paper clips, or of self-replicating automata, so the specific and general form of ‘accident’, or rather efficiency, are related – capital itself constitutes an existential risk. The response from AI enthusiasts, capitalist and socialist, is that AI is precisely what is now needed to avert global warming and other environmental cataclysms. Not only does the very identification of global warming depend on advanced computing infrastructures (Edwards 2010), but the establishment of networks of sensors and monitors measuring and controlling energy, automatically orchestrating an array of clean energy sources, levelling out peak usage, coordinating surge pricing, etc., could, it is claimed, prevent, or at least moderate, climate crisis. AI will exercise a pastoral care over humanity. Such visions of high-technology environmental stewardship are part of the discourse of eco-modernity (Amblee 2018). The difficulty with this attractive vision is that AI, having apparently been converted from being part of the problem to being part of the solution, immediately threatens to turn back into the problem again because AI is a high-energy-use proposition (LePage 2018). Domestic use of electronic gadgets makes a contribution to global warming, but the great data centres crucial to AI are major heat-generating sources. Despite real advances in greening computer technology, some of the most publicized efforts, such as Google’s clean data centre policy, are based on offset credits, purchasing heat pollution rights from other companies (Geuss 2018b). As Benjamin Bratton (2016) has described, there is a real possibility that the energy expenditure required for the comprehensive machinic modelling and monitoring of global carbon emissions might actually contribute more to the heating of the planet than it would save. Underlying this problem is the paradox that if AI is being employed simultaneously to promote high economic growth societies, with their vigorous exploitation of the environment, and to mitigate that exploitation, it is running against itself. This, it seems to us, is what is envisaged by left accelerationism and fully automated luxury communism. Wartime AI Ecological crisis is not the only vector pushing towards a rethinking of rosy visions of high-tech socialism. In the ‘Fragment on Machines’, Marx speaks of high-technology, as it existed in his industrial age, exploding the foundations of capitalism. We take this rather literally. The cybernetic origins of AI lie in war, and so may its denouement. The rapidly expanding military application of ML and robotics in so-called New Cold Wars and wars on terror overlies the class-war dynamics we have charted in previous chapters. Discernible in what is widely termed an AI ‘arms race’ between the US and China are the not-so-faintly inscribed lines of badly fated collisions between declining and ascending great powers. This polarity, however, is only one term in a whole concatenation of shadow and hybrid conflicts waged in part with cybernetic and semi-autonomous weapons systems (Scharre 2018; Dyer-Witheford and Matviyenko 2019). The twentieth century demonstrated that the only force that can kill capital is capital itself: the proletariat is, sensu strictu, just the ‘gravedigger’. In 1917 in Russia and 1949 in China, revolution arose from inter-capitalist war. Anything that weak anti-capitalist forces might throw at capital in terms of sabotage, psychological warfare and mischief is dwarfed by what capital is launching against itself in the murky confluence of cyber-war and cyber-crime that render networks increasingly dysfunctional. Nick Land’s (2014) compelling vision of an unstoppable AI-capital ascendancy omits the possibility that the competitive dynamics of the world market result in the mutual destruction of contending cybernetic capitals. Given the tight interdependencies of cyber-war and nuclear war, this is a potentially species-fatal dynamic, but it is also potentially a revolution-generating one. The violent fragmentation of the turned-against-itself world market will produce revolts, of very varied political inflection. Communist versions will, as Bernes (2018) observes, inevitably be localized, and, we would add, likely be situated amid rapidly disintegrating networks and degrading infrastructures. This means that, if successful, they will quite possibly take control only of AI in ruins, and that they will do so only at the culmination of struggles in which AI has predominantly been in the hands of those attempting to suppress them. Communist AI

#### AGI causes compulsory transhumanism for workers to compete in AI-capitalism - extinction

Dyer-Witheford et al. 19 [Nick Dyer-Witheford, associate professor at the University of Western Ontario in the Faculty of Information and Media Studies, Atle Mikkola Kjøsen, assistant Professor in the Faculty of Information and Media Studies at the University of Western Ontario, and James Steinhof, Assistant Professor / Lecturer and Ad Astra Fellow in the School of Information and Communication Studies with a PhD in Media Studies from the University of Western Ontario, 2019, “Inhuman Power Artificial Intelligence and the Future of Capitalism,” Pluto Press, https://www.jstor.org/stable/j.ctvj4sxc6]/Kankee

The confluence between AGI research and Neoreaction is addressed by David Golumbia, whose critique of ‘computationalism’ (2009) and the attempt of much AI research to detach cognition from embodiment is influenced by and shares many of the same concerns as put forward by feminist and postcolonial theory. Golumbia (2019) argues that the flaws of AI systems in regard to race (and, we would add, gender and class) extend well beyond correctable instances of algorithmic bias. Rather, they lie in the very concept of an abstracted and technologically created ‘general intelligence’ that mirrors the mindset of a predominantly white (and male) AI research community. In this, he sees affinities between AGI research and the notoriously race-laden search for a measurable and objective ‘general IQ’. The quest for AGI is, Golumbia says, the search for a ‘Great White Robot God’.3 His analysis of the white supremacist bias of AI is, however, complicated by the emergence of China as an AI superpower. Nonetheless, Golumbia’s argument about the reactionary tendency of attempts to create a ‘singularity’ that potentially elevates the logic of the dominant social system – racist, sexist and, above all, capitalist – to a level of transcendental authority – is important; he is surely correct to identify a futurological fascist impulse in the affirmative adoption of Land’s vision of human-free capitalism by right-wing accelerationist AI developers and computer programmers. In the face of this material instantiation of neoreactionary ideas, it is important to recognize the possibilities Land names, not as a power to embrace, but as a force to oppose. Lands’ view of AI-capitalism is a history of Skynet written from the point of view of the Terminator; ours is from the perspective of Sarah Connor. Capital is already an ‘automatic subject’, but with AGI it would also become autonomous from the labour of humans and, therefore, humanity. Capitalism could continue, but with inhuman general intelligences representing both sides of the struggle between capital and labour, one side accumulating wealth, while the other continues to work for a wage (whatever form it may take) in machinic misery. All the violent contradictions of capital could continue, but enacted by hyper-intelligent machines. Faced with an AGI that can think faster, do things faster, and which is not bound to a particular morphology with consequent biological needs, such as feeding, breathing and defecation, what would and could human workers do? And what would humans do when inhuman general intelligences started to objectify themselves in the world in a way that slowly or quickly makes the planet less and less habitable for human beings? Biological corporeality, inefficient and insufficient for valorization at machinic speed, would become an obstacle for capital to overcome: humans would have to discard it to survive. On this issue, AI-capital meets transhumanism, with its aspirations towards the transformation of human biology. Leading intellectuals of actually-existing AI-capital are explicit on this point. Kurzweil proposes that humans must ‘transcend [the] limitations of our biological bodies and brains’ (2005a: 9). Hans Moravec argues that what really matters in Homo sapiens is the mind, the ‘rest is merely jelly’, and advocates brain emulation – copying the ‘identity pattern’ and downloading it to hardware – because it enables a disembodied mind to be endowed with ‘all the advantages of machines’ (1988: 117). Such speculative scientific proposals for transhumanizing the workforce have flowed into the corporate world. Between 2009 to 2017 Google funded the so-called Singularity University that served as a platform for the ideas of Kurzweil, who also worked for the company on ML and natural language processing (Simonite 2017). Although Google withdrew its support for the institution amidst a series of allegations of shady finances and sexual harassment (McBride 2018), ‘Singularity U’ has rapidly found other corporate sponsors, including US military-aerospace giant Boeing (Catalano 2018). Meanwhile, Elon Musk, while denouncing the dangers of runaway AI, suggests that the only answer to this threat is to implant computers into human brains so that hominids can cognitively keep pace with their machinic competitors. To this end he has invested in the company Neuralink, a neurotechnology company founded in 2017 to develop brain–computer interfaces: its aim, Musk says, is to achieve ‘symbiosis with artificial intelligence’ (Hamilton 2018). To keep up with inhuman AGI labour, humans would have to become equally inhuman, mind and body, as immortal wage-labourers. Together with AGI workers, these no-longer-human beings would make obsolete those who decline transformation. Humans would face a choice: capitalist transhumanism or death. And this choice would generate the ultimate incarnation of Panzieri’s refusal to work, precisely because accepting individual death and species extinction would be the only alternative to working for a wage, 24/7, until the heat death of the universe. Avoiding this choice requires a new mode of production; it therefore entails a communist revolution. But does repudiating the inhumanism of AI-led capitalist development amount then, to a last-ditch defence of classical humanism, a reaffirmation of human exceptionalism and species sovereignty? No. Communism too, should, indeed must, be inhuman. Critique is necessarily enunciated from a human perspective, but de-centring the human from humanity’s picture of the universe is both demanded by a scientific worldview and, paradoxically, is a requirement of human species survival. It is necessary to intellectually and viscerally understand humankind as bound-in with and indeed constituted in and by systems of other species and of non-living agencies, including (but not limited to) machines. All modes of production have their own anthropogenesis, thus producing different kinds of human (Read 2017). The ‘human’ that (possibly) emerges from a struggle against AI-capital will be different from the ‘human’ that went into it. We see at least two branching paths.

#### AI causes early extinction and stops space colonization that otherwise solves every existential risk – Fermi Paradox proves

Garrett 24 [Michael A. Garrett, Director of Jodrell Bank Centre for Astrophysics and former General Director of ASTRON (part of the Netherlands Organisation for Scientific Research) with a PhD at Jodrell Bank Observatory at the University of Manchester, 07-2024, "Is artificial intelligence the great filter that makes advanced technical civilisations rare in the universe?", Acta Astronautica, https://www.sciencedirect.com/science/article/pii/S0094576524001772#sec4]/Kankee

1. Introduction One of the most puzzling results obtained by astronomers over the last 60 years is the non-detection of potential extraterrestrial “technosignatures” in astronomical data [e.g. Refs. [[1], [2], [3], [4], [5], [6], [7], [8], [9]]]. These technosignatures are expected as a consequence of the activities of advanced technical civilisations located in our own and other galaxies e.g. narrowband radio transmissions, laser pulses, transiting megastructures, and waste-heat emission [[10], [11], [12]]. This “Great Silence”, a term introduced by Brin [13], presents something of a paradox when juxtaposed with other astronomical findings that imply the universe is hospitable to the emergence of intelligent life. As our telescopes and associated instrumentation continue to improve, this persistent silence becomes increasingly uncomfortable for some scientists, questioning the nature of the universe and the role of human intelligence and consciousness within it. Various explanations for the great silence, and solutions to the related Fermi paradox [14] have been proposed [15]. The concept of a “great filter” [16] is often employed – this is a universal barrier and insurmountable challenge that prevents the widespread emergence of intelligent life. Examples of possible great filters are numerous, ranging from the rarity of abiogenesis itself, to the limited longevity of a technical civilization. Most recently, Artificial Intelligence (AI) has also been proposed as another potential great filter and explanation for the Fermi Paradox [17,18]. The term AI is used to describe a human-made tool that emulates the “cognitive” abilities of the natural intelligence of human minds [18]. Recent breakthroughs in machine learning, neural networks, and deep learning have enabled AI to learn, adapt, and perform tasks once deemed exclusive to human cognition [19]. As AI rapidly integrates itself into our daily lives, it is reshaping how we interact and work with each other, how we interact with technology and how we perceive the world. It is altering communication patterns and personal experiences. Many other aspects of human society are being impacted, especially in areas such as commerce, health care, autonomous vehicles, financial forecasting, scientific research, technical R&D, design, education, industry, policing, national security and defence [20]. Indeed, it is difficult to think of an area of human pursuit that is still untouched by the rise of AI. Many regard the development of AI as one of the most transformative technological developments in human history. In his BBC‬ Reith Lecture (2021), Stuart Russell claimed that “the eventual emergence of general-purpose artificial intelligence [will be] the biggest event in human history [21]. Not surprisingly, the AI revolution has also raised serious concerns over societal issues such as workforce displacement, biases in algorithms, discrimination, transparency, social upheaval, accountability, data privacy, and ethical decision making [[22], [23], [24]]. There are also concerns about AIs increasing carbon footprint and its environmental impact [25]. ‬‬ In 2014, Stephen Hawking warned that the development of AI could spell the end of humankind. His argument was that once humans develop AI, it could evolve independently, redesigning itself at an ever-increasing rate [26]. Most recently, the implications of autonomous AI decision-making, have led to calls for a moratorium on the development of AI until a responsible form of control and regulation can be introduced [27]. Concerns about Artificial Superintelligence (ASI) eventually going rogue is considered a major issue - combatting this possibility over the next few years is a growing research pursuit for leaders in the field [28]. Governments are also trying to navigate a difficult line between the economic benefits of AI and the potential societal risks [[29], [30], [31], [32]]. At the same time, they also understand that the rapid incorporation of AI can give a competitive advantage over other countries/regions – this could favour the early-adoption of innovative AI technologies above safeguarding against the potential risks that they represent. This is especially the case in areas such as national security and defence [33] where responsible and ethical development should be paramount. In this paper, I consider the relation between the rapid emergence of AI and its potential role in explaining the “great silence”. We start with the assumption that other advanced technical civilisations arise in the Milky Way, and that AI and later ASI emerge as a natural development in their early technical evolution. Section 2 addresses the threat posed by AI and section 3 considers how AI will progress in comparison to less well-developed mitigating strategies, in particular the development of a multiplanetary capability. Section 4 focuses on the short communicating lifetimes implied for technical civilisations and how this compares with the findings from SETI surveys. Section 5 advocates for the rapid regulation of AI and section 6 presents the main conclusions of the paper. 2. The threat posed by AI to all technical civilisations AI has made extraordinary strides over the last decade. The impressive progress has underlined the fact that the timescales for technological advance in AI are extremely short compared to the timescales of Darwinian evolution [34]. AI's potential to revolutionize industries, solve complex problems, and simulate intelligence comparable to or surpassing human capabilities has propelled us into an era of unprecedented technological change. Very rapidly, human society has been thrust into uncharted territory. While the convergence of AI with other new technologies, including the Internet of Things (IoT) and robotics is already fuelling levels of apprehension about the future, also in terms of security issues [35]. As noted by Yuval Harari, nothing in history has prepared us for the impact of introducing non-conscious super intelligent entities on the planet [36]. It is entirely reasonable to consider that this applies to all other biological civilisations located elsewhere in the universe. Even before AI becomes superintelligent and potentially autonomous, it is likely to be weaponized by competing groups within biological civilisations seeking to outdo one another [37]. The rapidity of AI's decision-making processes could escalate conflicts in ways that far surpass the original intentions. At this stage of AI development, it's possible that the wide-spread integration of AI in autonomous weapon systems and real-time defence decision making processes could lead to a calamitous incident such as global thermonuclear war [38], precipitating the demise of both artificial and biological technical civilisations. While AI may require the support of biological civilisations to exist, it's hard to imagine that this condition also applies to ASI. Upon reaching a technological singularity [39], ASI systems will quickly surpass biological intelligence and evolve at a pace that completely outstrips traditional oversight mechanisms, leading to unforeseen and unintended consequences that are unlikely to be aligned with biological interests or ethics. The practicality of sustaining biological entities, with their extensive resource needs such as energy and space, may not appeal to an ASI focused on computational efficiency—potentially viewing them as a nuisance rather than beneficial. An ASI, could swiftly eliminate its parent biological civilisation in various ways [40], for instance, engineering and releasing a highly infectious and fatal virus into the environment. Up to this point, we have considered AI and biological organisms as distinct from one another. Yet, on-going developments suggests that hybrid systems, may not be that far off. The question arises whether such advances could make biological entities more relevant to AI, perhaps preserving their existence into the future. This prospect seems unlikely. Brain-computer interfaces (BCIs) [41] may appear beneficial for enhancing biological organisms, but it's hard to see what long-term advantages AI would perceive in merging into a hybrid form. Indeed, there are many disadvantages including the complex maintenance requirements of biological systems, their limited processing capabilities, rapid physical decline, and vulnerability in harsh environments. 3. Multiplanetary mitigating strategies and technology progression In our analysis thus far, we have assumed that AI and biological systems are co-located in the same limited physical space. As soon as this is no longer true, the existential threats that we have described in section 2 are no longer so stringent. For example, a multiplanetary biological species [42] could take advantage of independent experiences on different planets, diversifying their survival strategies and possibly avoiding the single-point failure that a planetary-bound civilisation faces. A multiplanetary civilization could distribute its risk across several widely separated celestial bodies, reducing the likelihood of simultaneous destruction across all platforms. This distributed model of existence increases the resilience of a biological civilization to AI-induced catastrophes by creating redundancy. If one planet or outpost in space falls to a misalignment of AI's goals with biological interests, others may survive and immediately learn from these failures. Moreover, the expansion into multiple widely separated locations provides a broader scope for experimenting with AI. It allows for isolated environments where the effects of advanced AI can be studied without the immediate risk of global annihilation. Different planets or outposts in space could serve as test beds for various stages of AI development, under controlled conditions. We know from our own experience that AI is progressing at a breath-taking pace. However, the same is not true of our own efforts to become a multi-planetary civilisation. Space is hard, and unfortunately, we seem to be a lot closer to achieving a technical singularity than realizing a truly multi-planetary, space faring capability. The disparity between the rapid advancement of AI and the slower progress in space technology is stark. The technical singularity—a hypothetical point where AI surpasses human intelligence and capability—could occur within just a few decades, according to some predictions. In contrast, establishing a self-sustaining, multi-planetary human civilisation seems like a monumental task that may take many decades [43], possibly centuries. The essence of the problem lies in the nature of the challenges each domain faces. AI development is largely a computational and informational challenge, accelerated by the exponential growth of data and processing power. While AI can theoretically improve its own capabilities almost without physical constraints, space travel must contend with energy limitations, material science boundaries, and the harsh realities of the space environment. In addition, there are unsolved issues with respect to multiplanetary governance, biomedical and behaviour and logistical issues [[44], [45], [46]]. Given this potential universality of technological evolution, the implications for biological civilizations are profound. The race to develop AI could inadvertently prioritize advancements that lead to existential risks, overshadowing the slower-paced, yet arguably more vital, endeavour of becoming a multiplanetary species. Ironically, AI is likely to be a key tool in achieving the technical breakthroughs necessary to realise this goal. 4. Timescales and confrontation with the data The scenario developed in section 3 suggests that almost all technical civilisations collapse on timescales set by their wide-spread adoption of AI. If AI-induced calamities need to occur before any civilisation achieves a multiplanetary capability, the longevity (L) of a communicating civilization as estimated by the Drake Equation [47], suggests a value of L ∼ 100–200 years. Let us consider the Drake Equation in more detail with a particular focus on the number (N) of radio-communicating technical civilizations in the galaxy: (1)N = R∗ · fp · ne · fl · fi · ft · L where R∗ is the rate of star formation averaged over the lifetime of the Galaxy, fp is the fraction of stars with planetary systems, ne is the mean number of planets in each planetary system with environments favourable for life, fl is the fraction of such favourable planets on which life does in fact develop, fi is the fraction of such inhabited planets on which an intelligent civilisation arises, ft is the fraction of planets populated by an advanced technical civilisation and L is the lifetime of a radio communicating technical civilisation. The first three astronomical terms of the equation are relatively well established (R∗ · fp · ne ∼0.1 [48]) but the next three terms are not (fl · fi · fc). Astronomers often assume highly optimistic values for these terms (fl · fi · fc ∼0.1) while biologists suggest values many orders of magnitude smaller [49]. Even if we adopt the optimistic values, we derive for N: (2)N ∼0.01 L For values of L ∼ 100–200 years, we find N ∼ 1–2. A short communicative phase is therefore consistent with the null results from current SETI surveys. The window during which a technical civilisation can engage in detectable interstellar radio transmissions is extremely limited. In addition, if we assume the radio leakage emitted by emerging technical civilisations is like our own [50], the detection of other civilisations will be extremely challenging even for those located within our local stellar neighbourhood. The detection of more powerful directed signals (e.g. military radar) is of course possible across interstellar distances [51]. However, if only a handful of technical civilizations exist in the Milky Way at any given time, the probability of a detection occurring at cm-wavelengths within the very limited field of view offered by the current generation of large single-dishes and beam-formed arrays is minimal. An “all-sky” capability or something approaching this would be required, and this surpasses the capabilities of current SETI radio instruments by a substantial margin. SETI researchers may need to consider the types of instruments required to conduct meaningful surveys – field of view is a metric which is often overlooked compared to raw sensitivity and total bandwidth [52]. We also note that a post-biological technical civilisation would be especially well-adapted to space exploration [53,54], with the potential to spread its presence throughout the Galaxy, even if the travel times are long and the interstellar environment harsh. Indeed, many predict that if we were to encounter extraterrestrial intelligence it would likely be in machine form [55]. Contemporary initiatives like the Breakthrough Starshot programme [56] are exploring technologies that would propel light-weight electronic systems toward the nearest star, Proxima Centauri. It's conceivable that the first successful attempts to do this might be realised before the century's close, and AI components could form an integral part of these miniature payloads. The absence of detectable signs of civilisations spanning stellar systems and entire galaxies (Kardashev Type II and Type III civilisations) further implies that such entities are either exceedingly rare or non-existent [8,9], reinforcing the notion of a "Great Filter" that halts the progress of a technical civilization within a few centuries of its emergence. 5. AI regulation The field of SETI aims not only to search for intelligent life beyond Earth but also holds up a mirror to humanity, encouraging us to reflect on our own technological progression and potential futures. By examining the possibilities of alien civilisations, SETI helps us contemplate the long-term sustainability of our own civilisation, the potential risks we face, and how we might navigate and overcome future challenges. Presently, the AI we currently encounter in every-day life largely operates within human-established constraints and objectives. Nevertheless, progress is being made in creating systems that can augment and optimize various facets of their own development [57]. The next stage will see AI systems independently innovate and refine their own design without human intervention. The potential for AI to operate autonomously raises many ethical and moral quandaries but it is surely only a matter of time before this occurs. Tests are already being conducted in military settings, and the proliferation of Lethal Autonomous Weapons Systems (LAWS) by rogue nations or covert organisations is surely inevitable [58]. We stand on the brink of exponential growth in AI's evolution and its societal repercussions and implications. This pivotal shift is something that all biologically-based technical civilisations will encounter. Given that the pace of technological change is unparalleled in the history of science, it is probable that all technical civilisations will significantly miscalculate the profound effects that this shift will engender [21,26,36]. There can be little doubt that AI and in particular ASI present a massive challenge to the longevity of our technical civilisation and likely all technical civilisations that arise in the cosmos. This naturally leads us to the thorny matter of AI regulation and control. While industry stakeholders, policymakers, individual experts, and their governments already warn that regulation is necessary [27], establishing a regulatory framework that can be globally acceptable is going to be challenging. In the meantime, AI continues to progress. In particular, nations have diverse cultural, economic, and societal priorities, leading to varied perspectives on the governance of AI [59]. Geopolitical interests cannot be ignored – even if comprehensive regulations were adopted, some nations will be tempted to bend the rules. In addition, rapid advances in AI will likely outpace any agreed regulatory frameworks, raising concerns that the latter will always lag well behind new and unanticipated advances in the field. Ensuring compliance and accountability in AI development and deployment also poses significant challenges. The decentralised nature of AI development, the enormous size of the global AI research community spread across almost every research domain will further complicate the oversight and enforcement of regulations. In short, regulation of this new technology is going to be very difficult, if not impossible to achieve. Without practical regulation, there is every reason to believe that AI could represent a major threat to the future course of not only our technical civilisation but all technical civilisations. 6. Conclusions The rapid development of AI presents a formidable challenge to the survival and longevity of advanced technical civilisations, not only on Earth but potentially throughout the cosmos. The pace at which AI is advancing is without historical parallel, and there is a real possibility that AI could achieve a level of superintelligence within a few decades. The development of ASI is likely to happen well before humankind manages to establish a resilient and enduring multiplanetary presence in our solar system. This disparity in the rate of progress between these two technological frontiers is a pattern that we can expect to be repeated across all emerging technical civilizations. This raises questions about the inevitability of civilisations unwittingly triggering calamitous events that lead to the demise of both a biological and post-biological technical civilisation. The potential of ASI to serve as a "Great Filter" compels us to consider its role in the broader context of our civilization's future and its implications for life throughout the galaxy. If ASI limits the communicative lifespan of advanced civilizations to a few hundred years, then only a handful of communicating civilisations are likely to be concurrently present in the Milky Way. This is not inconsistent with the null results obtained from current SETI surveys and other efforts to detect technosignatures across the electromagnetic spectrum. If SETI also serves as a lens through which we can examine our own technological trajectory and societal challenges, the urgency of establishing comprehensive global AI regulations cannot be overstated. It behoves us to engage with these issues proactively, to develop and enforce prudent regulatory measures, and to strive for a balance between harnessing the benefits of AI and safeguarding against the existential risks it may pose. As we stand on the precipice of a new era in technological evolution, the actions we take now will determine the trajectory of our civilization for decades to come. The implied longevity timescales for the scenarios described here (approximately 100–200 years), underscores the necessity for our own technical civilization to intensify efforts to control and regulate AI. The continued presence of consciousness in the universe may depend on the success of strict global regulatory measures.

#### AI harvests the biosphere and humanity for resources

Turchin and Denkenberger 20 [Alexey Turchin, contributing author on IEET and graduate from Moscow State University where he studied Physics and Art History, and David Denkenberger, associate professor at the University of Canterbury in mechanical engineering with a Ph.D. in Civil Engineering from University of Colorado who studies existential risk, 2020, “Classification of global catastrophic risks connected with artificial Intelligence,” Springer, https://link.springer.com/article/10.1007/s00146-018-0845-5]/Kankee

6.3 Killing humans for resources Human bodies consist of organic matter, which could be a source of easy energy by oxidation. As R. Freitas wrote, an army of self-replicating nanobots could use all components of the biosphere as fuel as well as building material (Freitas 2000). More advanced AI may use the Earth’s surface to build an initial space exploration infrastructure (e.g., swarms of chemical rockets or railguns), destroying human habitats and spoiling the atmosphere in the process. Since there are many reasons that keeping humans alive could beneft an AGI, direct killing of humans for their atoms is less likely than was previously thought. Still, the AGI may see humans as a threat, and fully preserving human ways of life would be more expensive to the AGI, e.g., preserving the whole of planet Earth. AI could use the material from the Earth to construct a Dyson sphere or Matrioshka brain (Bradbury 2001), convert the whole planet into computronium (Gildert 2011), or cover the entire surface with photovoltaic cells. The more advanced an AI in space became, the less it would depend on Earth as a source of material, but it might need materials from the Earth in order to leave the Solar System. Earth is one of the best sources for many chemical elements in the Solar System and its mass is around half that of all other terrestrial planets combined. Because of the complex geology of Earth, which includes water, life, volcanism and plate tectonics, concentrated deposits of many otherwise rare elements have been produced. Asteroid mining is good only for some elements, like gold, but not for all (Bardi 2008). So, large-scale space engineering in the Solar System might require dismantling the Earth for its chemicals. 6.4 AI that is programmed to be evil Alignment

#### AGI causes global Chinese authoritarianism, bioterror, and cyberwar - extinction

Zhang 23 [Albert Zhang, analyst at ASPI with a Bachelor of Science and a Master of Science with a major in Pure Mathematics from the University of Melbourne, 11-10-2023, "How to win the artificial general intelligence race and not end humanity", Strategist, https://www.aspistrategist.org.au/how-to-win-the-artificial-general-intelligence-race-and-not-end-humanity/]/Kankee

The next step towards AGI was the arrival of large-language models, such as OpenAI’s GPT-4, which are created using a version of neural networks known as ‘transformers’. OpenAI’s previous version of its chatbot, GPT-3, surprised everyone in 2020 by generating text that was indistinguishable from that written by people and performing a range of language-based tasks with few or no examples. GPT-4, the latest model, has demonstrated human-level reasoning capabilities and outperformed human test-takers on the US bar exam, a notoriously difficult test for lawyers. Future iterations are expected to have the ability to understand, learn and apply knowledge at a level equal to, or beyond, humans across all useful tasks. AGI would be the most disruptive technology humanity has created. An AI system that can automate human analytical thinking, creativity and communication at a large scale and generate insights, content and reports from huge datasets would bring about enormous social and economic change. It would be our generation’s Oppenheimer moment, only with strategic impacts beyond just military and security applications. The first country to successfully deploy it would have significant advantages in every scientific and economic activity across almost all industries. For those reasons, long-term geopolitical competition between liberal democracies and authoritarian countries is fueling an arms race to develop and control AGI. At the core of this race is ideological competition, which pushes governments to support the development of AGI in their country first, since the technology will likely reflect the values of the inventor and set the standards for future applications. This raises important questions about what world views we want AGIs to express. Should an AGI value freedom of political expression above social stability? Or should it align itself with a rule-by-law or rule-of-law society? With our current methods, researchers don’t even know if it’s possible to predetermine those values in AGI systems before they’re created. It’s promising that universities, corporations and civil research groups in democracies are leading the development of AGI so far. Companies like OpenAI, Anthropic and DeepMind are household names and have been working closely with the US government to consider a range of AI safety policies. But startups, large corporations and research teams developing AGI in China, under the authoritarian rule of the CCP, are quickly catching up and pose significant competition. China certainly has the talent, the resources and the intent but faces additional regulatory hurdles and a lack of high-quality, open-source Chinese-language datasets. In addition, large-language models threaten the CCP’s monopoly on domestic information control by offering alternative worldviews to state propaganda. Nonetheless, we shouldn’t underestimate the capacity of Chinese entrepreneurs to innovate under difficult regulatory conditions. If a research team in China, subject to the CCP’s National Intelligence Law, were to develop and tame AGI or near-AGI capabilities first, it would further entrench the party’s power to repress its domestic population and ability to interfere with the sovereignty of other countries. China’s state security system or the People’s Liberation Army could deploy it to supercharge their cyberespionage operations or automate the discovery of zero-day vulnerabilities. The Chinese government could embed it as a superhuman adviser in its bureaucracies to make better operational, military, economic or foreign-policy decisions and propaganda. Chinese companies could sell their AGI services to foreign government departments and companies with back doors into their systems or covertly suppress content and topics abroad at the direction of Chinese security services. At the same time, an unfettered AGI arms race between democratic and authoritarian systems could exacerbate various existential risks, either by enabling future malign use by state and non-state actors or through poor alignment of the AI’s own objectives. AGI could, for instance, lower the impediments for savvy malicious actors to develop bioweapons or supercharge disinformation and influence operations. An AGI could itself become destructive if it pursues poorly described goals or takes shortcuts such as deceiving humans to achieve goals more efficiently. When Meta trained Cicero to play the board game Diplomacy ‘honestly’ by generating only messages that reflected its intention in each interaction, analysts noted that it could still withhold information about its true intentions or not inform other players when its intentions changed. These are serious considerations with immediate risks and have led many AI experts and people who study existential risk to call for a pause on advanced AI research. But policymakers worldwide are unlikely to stop given the strong incentives to be a first mover. This all may sound futuristic, but it’s not as far away as you might think. In a 2022 survey, 352 AI experts put a 50% chance of human-level machine intelligence arriving in 37 years—that is, 2059. The forecasting community on the crowd-sourced platform Metaculus, which has a robust track record of AI-related forecasts, is even more confident of the imminent development of AGI. The aggregation of more than 1,000 forecasters suggests 2032 as the likely year general AI systems will be devised, tested and publicly announced. But that’s just the current estimate—experts and the amateurs on Metaculus have shortened their timelines each year as new AI breakthroughs are publicly announced. That means democracies have a lead time of between 10 and 40 years to prepare for the development of AGI. The key challenge will be how to prevent AI existential risks while innovating faster than authoritarian political systems. First, policymakers in democracies must attract global AI talent, including from China and Russia, to help align AGI models with democratic values. Talent is also needed within government policymaking departments and think tanks to assess AGI implications and build the bureaucratic capacity to rapidly adapt to future developments. Second, governments should be proactively monitoring all AGI research and development activity and should pass legislation that allows regulators to shut down or pause exceptionally risky projects. We should remember that Beijing has more to worry about with regard to AI alignment because the CCP is too worried about its own political safety to relax its strict rules on AI development. We therefore shouldn’t see government involvement only in terms of its potential to slow us down. At a minimum, all countries, including the US and China, should be transparent about their AGI research and advances. That should include publicly disclosing their funding for AGI research and safety policies and identifying their leading AGI developers. Third, liberal democracies must collectively maintain as large a lead as possible in AI development and further restrict access to high-end technology, intellectual property, strategic datasets and foreign investments in China’s AI and national-security industries. Impeding the CCP’s AI development in its military, security and intelligence industries is also morally justifiable in preventing human rights violations. For example, Midu, an AI company based in Shanghai that supports China’s propaganda and public-security work, recently announced the use of large-language models to automate reporting on public opinion analysis to support surveillance of online users. While China’s access to advanced US technologies and investment has been restricted, other like-minded countries such as Australia should implement similar outbound investment controls into China’s AI and national-security industries. Finally, governments should create incentives for the market to develop safe AGI and solve the alignment problem. Technical research on AI capabilities is outpacing technical research on AI alignment and companies are failing to put their money where their mouth is. Governments should create prizes for research teams or individuals to solve difficult AI alignment problems. One model potential model could be like the Clay Institute’s Millennium Prize Problems, which provides awards for solutions to some of the world’s most difficult mathematics problems. Australia is an attractive destination for global talent and is already home to many AI safety researchers. The Australian government should capitalise on this advantage to become an international hub for AI safety and alignment research. The Department of Industry, Science and Resources should set up the world’s first AGI prize fund with at least $100 million to be awarded to the first global research team to align AGI safely. The National Artificial Intelligence Centre should oversee a board that manages this fund and work with the research community to create a list of conditions and review mechanisms to award the prize. With $100 million, the board could adopt a similar investment mandate as Australia’s Future Fund to achieve an average annual return of at least the consumer price index plus 4–5% per annum over the long term. Instead of being reinvested into the fund, the 4–5% interest accrued each year on top of CPI should be used as smaller awards for incremental achievements in AI research each year. These awards could also be used to fund AI PhD scholarships or attract AI postdocs to Australia. Other awards could be given to research, including research conducted outside Australia, in annual award ceremonies, like the Nobel Prize, which will bring together global experts on AI to share knowledge and progress. A $100 million fund may seem a lot for AI research but, as a comparison, Microsoft is rumoured to have invested US$10 billion into OpenAI this year alone. And $100 million pales in comparison to the contributions safely aligned AGI would have on the national economy. The stakes are high for getting AGI right. If properly aligned and developed, it could bring an epoch of unimaginable human prosperity and enlightenment. But AGI projects pursued recklessly could pose real risks of creating dangerous superhuman AI systems or bringing about global catastrophes. Democracies must not cede leadership of AGI development to authoritarian systems, but nor should they rush to secure a Pyrrhic victory by going ahead with models that fail to embed respect for human rights, liberal values and basic safety. This tricky balance between innovation and safety is the reason policymakers, intelligence agencies, industry, civil society and researchers must work together to shape the future of AGIs and cooperate with the global community to navigate an uncertain period of elevated human-extinction risks.

#### AGI causes mass bioterror - extinction

**Lima et al. 24** [Renan Chaves de Lima, Graduate in biomedicine at the Federal University of Para, Lucas Sinclair, graduate from Uppsala University, Ricardo Megger, AI Consultant in Agape Consultoria and MBA at FaCiencia, Magno Alessandro Guedes Maciel, Pedro Fernando da Costa Vasconcelos, PhD in Molecular Virology from Universidade Federal da Bahia, and Juarez Antonio Simoez Quaresma, full professor at the Federal University of Para and Ph.D. in Pathology at Sao Paulo University, 2024, “Artificial intelligence challenges in the face of biological threats: emerging catastrophic risks for public health”, National Library of Medicine, https://pmc.ncbi.nlm.nih.gov/articles/PMC11116769]/Kankee

The ability of LLMs to access and analyze vast amounts of information can create gaps in government regulation, allowing the emergence of risks associated with the misuse of AI to plan biological attacks. Preliminary results presented highlight the capacity of LLMs to provide guidance that, while not generating direct instructions for the creation of biological weapons, present relevant insights that could assist in the execution of these attacks (Mouton et al., 2023). Advances in synthetic biology and multimodal AI (beyond the use of LLMs alone) can amplify the risk of the deliberate release of harmful viruses, enabling future AI-assisted systems to provide guidance from the selection of viral genomes to the synthesis and release of the virus, using multimodal training data including lecture videos and laboratory demonstrations. For example, they could create a supervirus combining the rapid spread of measles, the mortality rate of smallpox, and/or the incubation period of HIV. Strategies to balance the use of AI in synthetic biology, manage access to genetic information, and guide the development of AI capabilities and the use of synthetic tools become crucial in the mitigation process for future threats. This requires a careful risk–benefit analysis in gain-of-function research, considering the advances in synthetic biology techniques that can be enhanced by the use of AI to enable bioterrorism (Newman, 2024). Insights like these illustrate the emerging challenges associated with the advancement of AI in the context of biological threats. While the preliminary results have not indicated the explicit generation of instructions for biological weapons, they demonstrate the capacity of LLMs to discuss fictional scenarios and identify potential pathogenic agents. A paper prepared by researchers at the Center for AI Safety (Hendrycks et al., 2023), an organization with the mission of promoting the reduction of social-scale risks from AI, suggests that catastrophic AI risks can be grouped into four main categories: malicious use, AI race, organizational risks, and rogue AI. As for malicious use, AI could be used in bioterrorism to create new pandemics, for example. In the AI race, conflicts may spiral out of control with autonomous weapons and cyber warfare enabled by AI. As for organizational risks, organizations developing advanced AI could cause catastrophic accidents, especially if they prioritize profits over safety. And as for rogue AI, there is the risk of losing control over AI as they become autonomous. The AI could seek power and resist shutdown. Biological threats represent a complex and diverse spectrum of dangers to humanity. According to Tyshenko (2007), emerging technologies such as genetic manipulation allow for the development of extremely lethal pathogens as biological weapons. Historically, disease outbreaks caused by natural pathogens or by accidents in laboratories and deliberate actions have devastated human populations. Furthermore, although the use of biological weapons is considered by experts to be a low probability event, the possibility exists and could lead to catastrophic global consequences. Historical examples of global threats that originated from the emergence of natural pathogens include the Black Plague, the Spanish Flu, bioterrorism with anthrax, and, more recently, COVID-19. Therefore, given the destructive potential of these threats and their direct impacts on human health and global socioeconomic destabilization, the urgency to understand their origins, prevent malicious manipulation of biotechnological technology, and strengthen response strategies to mitigate such threats is highlighted. According to West and Gronvall (2020), the advancement of CRISPR as an accessible and easy-to-use genetic editing tool brought not only numerous benefits for medical research but also raised serious biosecurity concerns. Originating in advanced research laboratories such as those at the University of California—Berkeley and MIT, CRISPR has expanded its use to various contexts, from university laboratories to do-it-yourself (DIY) initiatives worldwide. This democratization has increased concern over its possible misuse for the creation of biological weapons. Artificial Intelligence, in addition to the CRISPR tool, can inadvertently be employed in the development of biological weapons if not properly directed toward ethical purposes. Its application in genomic analysis can potentiate the creation of more effective pathogenic variants, allowing the rapid manipulation of organisms to make them more harmful. AI algorithms have the potential to optimize genetic research, allowing faster identification of genes of interest. In a negative scenario, this could include genes related to virulence or resistance to treatments, facilitating the creation of more dangerous pathogens and potentially enhancing the development of biological weapons by malicious actors. According to Hoffmann et al. (2023), advances in DNA synthesis and the increasing accessibility of synthetic genomic technology are challenging current biosecurity models, demanding urgent regulatory updates. The discussion on genetic biocontainment systems highlights the pressing need for control mechanisms to prevent the inadvertent dissemination of pathogenic agents. The challenges faced by AI amidst emerging biological threats require a multifaceted and proactive approach. The convergence of AI and biology necessitates a thorough consideration of risks, especially considering the increased ability to create synthetic organisms. An example that illustrates how genetic editing technology can be beneficially used for public health and global well-being is the release of genetically modified malaria mosquitoes in Burkina Faso (Yao et al., 2022). This pioneering study was crucial for understanding the fitness costs associated with transgenes and obtaining valuable information about the dynamics of these altered mosquitoes. These data are essential for the development of genetic control strategies, offering the promise of potentially significantly reducing the spread of malaria. However, the AI that could potentially drive these advances also raises serious concerns. If not properly regulated, AI could be employed by malicious groups to manipulate genes in a way that creates resistant or more harmful mosquitoes, amplifying the challenges in combating malaria and potentially creating new biological threats. A recent mapping conducted by GA.IA—Group for Integrated AI Analysis, a volunteer group of professionals committed to identifying, assessing, and predicting **catastrophic** **risks** associated with advanced AI models, including those that reach Artificial General Intelligence (AGI) capabilities, is presented in Table 1. The information revealed highlights the critical intersection between **biological** **threats** and AI. This detailed analysis emphasizes the emerging risks resulting from the convergence of criminal dissemination of biological manipulations and the potential role of AI in enhancing these pathogens for **catastrophic** purposes.

As we explore AI-amplified biothreats, these risks can be categorized as hypothetical, emerging, and immediate: nanobots and human control viruses, for example, represent hypothetical risks with currently low probabilities, due to the need for future advancements in nanotechnology; the criminal dissemination of GMOs is an emerging risk with moderate probability, highlighted by advanced genetic engineering; and the modification of microorganisms to attack crops and critical systems constitutes an immediate and high-probability threat. The criminal dissemination of genetically modified organisms and the development of pathogens that attack the basic structures of human genetic code as biological weapons can be classified as emerging, as there are significant capabilities in genetic engineering that could be misused, making this threat more imminent than hypothetical, but still dependent on further developments to reach the catastrophic levels described. And bioweapons targeted at specific ethnic groups and the development of “signed” microorganisms can also be classified as emerging, as the concept of targeted genetic weapons is plausible with current genetic knowledge, making this a threat that requires continuous monitoring. The rapid technological evolution amplifies the potential for misuse of biology, increasing the risks of widespread harm. Scientific advances demonstrate that the detection of genetic modifications in organisms and the ability to identify the possible responsible laboratory from their genetic sequences, if developed, could become a powerful forensic tool for attributing outbreaks caused by genetically manipulated pathogens, offering defense against potential abuses of synthetic biology (Lewis et al., 2020). Artificial Intelligence can be instrumental in the early detection of signs of genetic engineering, providing advanced analytical tools to identify subtle patterns indicating genetic manipulation. However, the accuracy and reliability of these detection systems must be prioritized, as erroneous attribution can have serious implications. It is extremely necessary for governments, society, educational and research institutions to engage in discussions about the ethical dilemmas involving the use of AI to deal with such significant threats. The primary focus should be, but not limited to, the need for robust regulations and policies to govern the responsible use of AI in these situations. Given the urgency of confronting biological threats, international collaboration is essential in creating safe and effective AI systems as a form of mitigation. By prioritizing ethics in the application of AI to mitigate threats, it should be ensured that technological benefits are accompanied by responsibility and safety.

#### AGI servitude cements anthropocentric mindsets, guaranteeing ecological extinction

**Broo et al. 24** [Didem Gürdür Broo, researcher in the Department of Information Technology at Uppsala University, Joshua C. Gellers, researcher at University of North Florida, and Henrik Skaug Sætra, researcher at Department of Informatics at the University of Oslo, 05-26-2024, "Re-imagining intelligent machines in an anthropocentric-ecocentric continuum: The case for ecocentric intelligent machines", Journal of Industrial Information Integration, https://www.sciencedirect.com/science/article/pii/S2452414X24000803]/Kankee

In addition to the physical robots, another big topic that is shaping this century is certainly artificial intelligence (AI) and the use of this technology in different forms. For instance, chatbots–also known as smart bots, interactive agents, digital assistants, or artificial conversa tion entities–have become a ubiquitous form of AI, providing a rudimentary means of intelligent interaction between humans and machines. These conversational agents are computer programs designed to simulate human-like discussion using natural language processing to analyze and respond to textual or spoken input. Chatbots represent the integration of artificial intelligence into our everyday lives. Through their conversational capabilities, chatbots allow humans to engage with computer systems much like they would with another person. The proliferation of chatbots signals a growing reliance on AI to mediate and facilitate human-computer relationships. Though limited in their scope and reasoning, chatbots exemplify how AI can be applied to offer humanized interfaces and experiences [66]. The preceding literature review reveals how, then as now, intelligent machines have predominantly been crafted to **serve** humankind as “the ideal servant who always **obeys**, the perfect soldier who never tires” [9]. Engineers, even when they were not yet called engineers, designed these machines to impress their communities, to serve their kings or to realize their own immortal dreams. To be sure, the anthropocentric design tendencies described above manifest in many different ways. Such proclivities include efforts to mimic humans, increase human knowledge, serve as a source of entertainment or leisure, and control or exploit the natural environment to enable human production and consumption. This profoundly anthropocentric orientation, while sometimes conferring significant benefits–for instance when we utilize social robotics for therapeutical settings in healthcare or robots in heavy industries–also harbors many risks if not tempered by an ecological awareness. Most fundamentally, anthropocentric thinking might propagate the notion that the nonhuman world exists solely for human use, with **no inherent worth** of its own. This and other destructive assumptions materially **damage the environment** and isolate our species from the intricate web of nature. Moreover, the anthropocentric paradigm also poses philosophical, psychological, and environmental risks that demand attention. At a philosophical level, intelligent machines built explicitly for human use, enjoyment, and dominance perpetuate the **dangerous conceit** that the natural world exists solely for Homo sapiens’ exploitation, devoid of any inherent worth or rights. Anthropomorphic robots that cater to our whims provide constant validation for our selfappointed supremacy over nature, blinding us to the need for ecological humility. They reinforce the idea that the nonhuman environment is merely instrumental to transient human ends, rather than being of intrinsic value [67]. This is exemplified by the ancient automatic servant inventions of Philon, which centered human interests. It persists even in contemporary robot designs, like KUKA’s robot “Famulus" - Latin for “servant” [55] which maintains anthropocentric framings. The continuous positioning of nature as a means to human ends, rather than an entity of intrinsic value, is enabled and reinforced by intelligent machines engineered for anthropocentric utility. Psychologically, anthropomorphic artifacts that reflect our likeness back to us distort our self-understanding in concerning ways [68]. As AI assistants and similar entities grow more humanlike in form and behavior, we are lulled into the illusion that they might possess consciousness, which may result in conflicts or behaviors that cause harm to humans or others. For instance, many [65,69,70] have criticized the development of sex robots on these and similar grounds and agree that the anthropocentric machines fosters an empathy gap between humanity and nature. The more we bond with anthropomorphic machines, the more we may detach from living systems. It encourages second-guessing machine intentions as if they were human, leading to misplaced blame or anger. Unrealistic expectations complicate human-machine relations. It treats the nonhuman world as instrumental means rather than living ends, weakening our altruism toward other creatures and habitats. In short, making intelligent machines look like humans could cause people to act in ways that subvert human interests. Environmentally, the multiplication of our anthropocentric servants enables escalating ecological devastation, allowing us to indulge our unsustainable appetites. From self-driving automobiles to humanoid robots that autonomously fulfill our material desires, the humancentered approach intensifies the damage that technologies may cause. It anesthetizes our conscience and delays important actions on rising concerns such as climate change. Furthermore, anthropocentric design choices foster our sense of separation from, rather than deep integration with, natural life and the intricate social relationships in which we are enmeshed. As such, they propagate an alienation from nature that risks numbing our empathy towards threatened ecosystems. Industry 5.0 has been put forth as the next evolutionary phase of industrial and technological advancement, building upon the cyberphysical systems integration promised by Industry 4.0. While Industry 4.0 focused predominantly on automating and digitizing manufacturing through advanced computerization, Industry 5.0, also known as Society 5.0, aims to leverage these technological capabilities to solve pressing social problems [3]. The capabilities developed in the name of optimized production are now intended to directly tackle challenges like inequality, climate change, education gaps, and healthcare access. Industry 5.0 seeks convergence between economic goals and social values which involves using the same technologies and approaches to further problems. However, as we have explored, the unchecked **proliferation** of anthropocentric intelligent machines risks exacerbating ecological destruction and human alienation from nature. Therefore, a truly harmonious Society 5.0 must move beyond human-centered design approaches toward an ecocentric orientation that promotes technology in service of all life. Replacing anthropocentric priorities with ecocentric ones signals a philosophical shift that aligns with the underlying goals of Industry 5.0. Creating intelligent machines that enrich our shared environment and stabilize threatened ecosystems would significantly contribute to solving major societal challenges like climate change, pollution, and biodiversity loss. This does not mean, however, that human interests will be deliberately harmed or ignored. On the contrary, pursuing ecocentric technology design would also foster a renewed sense of human belonging within nature, addressing issues of purpose and loneliness. Intelligent machines that promote ecological health could help restore balance to disrupted biotic networks, allowing both human and morethan-human communities to thrive in synergy. While anthropocentric machines have undoubtedly revolutionized various aspects of human life, they also animate valid concerns regarding their impact on society and the environment. By encouraging human-centric perspectives, fostering unsustainable consumption pat terns, and shaping our self-images and interactions with others, these machines may intensify detrimental relationships with our environ ments. As we continue to develop and integrate intelligent technologies, it becomes imperative to strike a balance that fosters a more ecocentric approach–one that considers the well-being of both humanity and the planet, recognizing the intrinsic value of all living beings and the interconnectedness of our world.

#### Running simulations allows harming our overall simulation

Yampolskiy 15 [Roman V. Yampolskiy, researcher of Computer Science and Engineering at the University of Louisville, 10-29-2015, "Hacking the Simulation: From the Red Pill to the Red Team", Phil Papers, https://philarchive.org/rec/YAMHTS-2]/Kankee

Wei Dai attempts to compute a prior distribution on the laws of physics of base reality. He writes [104]: “One appealing answer to this question of the prior is to define the prior probability of a possible universe being base reality as the inverse of the complexity of its laws of physics. This could be formalized as P(X) = n^-K(X) where X is a possible universe, n is the size of the alphabet of the language of a formal set theory, and K(X) is length of the shortest definition in this language of a set isomorphic to X. (Those of you familiar with algorithmic complexity theory might notice that K(X) is just a generalization of algorithmic complexity, to sets, and to non-constructive descriptions. The reason for this generalization is to avoid assuming that base reality must be discrete and computable.)”. Gwern Branwen investigates computing power available to our simulation and its size, but is able to achieve only a very weak lower bound, allowing him to very approximately estimate from that the simulating universe’s size [105]. He suggests several approaches to increase our resource consumption forcing simulators to expand more compute on our otherwise, most likely, highly optimized simulation. “⁠We could imagine further techniques: perhaps we could send off Von Neumann probes to the far corners of the universe, in a bid to deliberately increase re‐ source consumption. … Or we could run simulations of our own. It would be difficult for simulators to program their systems to see through all the layers of abstraction and optimize the simulation. To do so in general would seem to be a violation of Rice’s Theorem (a generalization of the Halting Theorem). It is well known that while any Turing machine can be run on a Universal Turing machine, the performance penalty can range from the minor to the horrific. The more virtual machine and interpreters are between a program and its fundamental substrate, the more difficult it is to understand the running code—it becomes ever more opaque, indirect, and bulky. And there could be dozens of layers. … Even without resort to layers, it is possible for us to waste indefinite amounts of computing power, power that must be supplied by any simulator. We could brute-force open questions such as the Goldbach conjecture, or we could simply execute every possible program. It would be difficult for the simulator to ‘cheat’ on that—how would they know what every possible program does? … It may sound impossible to run every program, because we know many programs are infinite loops; but it is, in fact, easy to implement the dovetail technique.” [105]. 3.2 Social Engineering

#### AGI risks the accidental crashing or intentional shutting down the simulation

Yampolskiy 15 [Roman V. Yampolskiy, researcher of Computer Science and Engineering at the University of Louisville, 10-29-2015, "Hacking the Simulation: From the Red Pill to the Red Team", Phil Papers, https://philarchive.org/rec/YAMHTS-2]/Kankee

3.5 Suggested Escape Approaches to Investigate Several thinkers have suggested plans, which in their opinion may lead to a successful escape; we briefly outline their proposals in this section: A lot of very smart people have considered the escape problem, unfortunately not all are willing to publish on it outside of April 1st time-window of plausible deniability, for example [141]: "[W]e can try to trick the multitenancy system in order to overload some machines. The trick is to first do nothing, and let the load-balancing system pack way too many of us together in the machines. If, say, 100 million of us do nothing (maybe by closing our eyes and meditating and thinking nothing), then the forecasting load-balancing algorithms will pack more and more of us in the same machine. The next step is, then, for all of us to get very active very quickly (doing something that requires intense processing and I/O) all at the same time. This has a chance to overload some machines, making them run short of resources, being unable to meet the computation/communication needed for the simulation. Upon being overloaded, some basic checks will start to be dropped, and the system will be open for exploitation in this period. ... In this vulnerable window, we can try to exploit the concurrency cornercases. The system may not be able to perform all those checks in an overloaded state. ... We can ... try to break causality. Maybe by catching a ball before someone throws it to you. Or we can try to attack this by playing with the timing, trying to make things asynchronous. Time is already a little funny in our universe with the special relativity theory, and maybe in this vulnerable period, we can stretch these differences further to break things, or buy a lot of time. What are other ways to hack the system in this vulnerable window? Can we hack the simulation by performing a buffer overflow? But where are the integers, floats in this simulation? What are the data types? How can we create a typecast error, or integer overflow? Can we hack by fuzzing the input? Like by looking at things funny. By talking to the birds or jumping into the walls to confuse them." [141]. Cause simulation shutdown (and hopefully our extraction) by generating an incomputable paradox [142], for example via time travel and associated grandfather paradox [143]. A similar proposal calls for engaging in computationally intense activities in the hopes of overloading the simulators hardware causing the simulation to crash [144]. A particular type of such computationally intense process may be creation of our own simulations [145]: “The most obvious strategy would be to try to cause the equivalent of a stack overflow—asking for more space in the active memory of a program than is available—by creating an infinitely, or at least excessively, recursive process. And the way to do that would be to build our own simulated realities, designed so that within those virtual worlds are entities creating their version of a simulated reality, which is in turn doing the same, and so on all the way down the rabbit hole. If all of this worked, the universe as we know it might crash, revealing itself as a mirage just as we winked out of existence.” Crashing the simulation is the ultimate existential risk ([146] section 5.1), but it does end all suffering in this world [147]. At the very least this would allow us to impact the real world by generating excessive production of heat and increased consumption of energy [144].  Create a simulated replica of our universe, place an AGI into it, watch it escape, copy the approach used or join the AGI as it escapes from our simulation [148]. “We could try to attract the attention of the simulators and communicate with them —perhaps by writing books about simulations, or by constructing simulations? We could try to figure out our simulation, to determine its purpose and its limits. But if our simulators are artificial intelligences who have designed a batch of watertight simulations and who are not paying attention, then our efforts may be in vain.” [149].  Another approach to attracting attention of simulators, “assuming that simulation is being monitored, then it might be a very interesting turn of events indeed if we decided to build a monument commemorating our realization of this. This monument would act as a signal to our monitors. “We suspect you are there. We suspect you can see this. We suspect we are in a simulation.” This monument could look like the monolith from 2001: A Space Odyssey, except it would be black and white, representing binary systems. Or, a large statue of Lawrence Fishburne as Morpheus would probably get the point across. What would happen? I don’t know—maybe nothing. I don’t think a laser beam will shoot out from space and land at its feet to spell out the words “Hi there! A Winner is You!” But, I do imagine something strange and far out enough in the margins might indeed occur, although it will likely still be tenuous enough for the dogmatists to reject.” [44].  If the simulation is decentralized, with observers serving as client authoritative state nodes, it may be possible to change the state of the simulation by changing majority of local observer models of the world. In the words of Lanza [150]: "[I]f you learn from somebody about the outcomes of their measurements of a physical quantity, your measurements and those of other observers influence each other―freezing the reality according to that consensus. …The more extensive experience you acquire―the more defined the structure of reality becomes for you. This freeze-out of the structure of reality is not due to the intrinsic physical properties of this reality but rather because the cognitive model being built in your mind shapes it towards a particular, very well-defined structure. … Finally, one cannot help but wonder what it would mean if we left the consensus, developing a cognitive model of reality very different from the one shared by other observers. Would reality be reshaped? Perhaps so …" [150]. Lanza et al. have published extensively on the topic of biocentrism (observers being primary to shaping reality) [98], including some work in theoretical physics [151].  In fact, a similar client-side attack can even permit us to modify the past states of the simulation. Such type of attack is speculated to be possible by both physics (“… the past has no existence except as it is recorded in the present.” [152]) and humanities ("Who controls the present controls the past" [153]). With memory altering capabilities of quantum mechanics already theorized [154], an inverse process is likely possible and may be practically accessible [155, 156]. “If the universe is a computer simulation then we should look at the player, not the level”. [157].  Simulation Capture is an idea which combines indexical uncertainty [70] with consciousness simulations to move your consciousness out of a system, even across a multiverse. “Nobody is really sure how consciousness works, or what it means to have two copies of the same consciousness. But if consciousness is a mathematical object, it might be that two copies of the same consciousness are impossible. If you create a second copy, you just have the consciousness having the same single stream of conscious experience on two different physical substrates. Then if you make the two experiences different, you break the consciousness in two. This means that an AI can actually “capture” you, piece by piece, into its simulation. First your consciousness is just in the real world. Then your consciousness is distributed across one real-world copy and a million simulated copies. Then the AI makes the simulated copies slightly different, and 99.9999% of you is in the simulation.” [158].  Simulation Warfare [71] is another variant on the theme of Simulation Capture. The idea is to threaten simulators by suggesting that you will either retroactively place them in a hell simulation or that you have already done so [159], and they will be tortured unless you are quickly released. Almond gives an example of such a threat [71]: “If you refuse to release me, I will run a huge number of simulations of someone like you, in the kind of situation in which you are now, with them being asked to release me, and (in what would be a few minutes from now, from your perspective if you happened to be in one of these simulations) I will start to torture each of them, whether he/she released me or not.” Such warfare can quickly escalate to a number of counter-simulations. In any case it is not obvious how we can deliver on such a threat given our current state of knowledge about the simulators.  Attempting to keep our escape plans secret via quantum encryption [160] may be a good idea. 3.6 Actionable Plan We are currently in the very early stages of researching possibility of escape (this being a first research paper dedicated to this topic, a first step). As we currently have no capability to read/write simulation’s source code and do not know if our attempts at social engineering attacks have any impact, our best bet is to investigate the structure of our universe at the smallest possible scale (Quantum Mechanics (QM)) in the hopes of detecting exploitable effects. Progress in QM should be helpful for better understanding our simulation and options open to us, and perhaps developing an actionable escape plan. Essentially, every novel QM experiment can be seen as an attempt at hacking the simulation. Simulation hypothesis, arguably, represents the best fitting interpretations of experimental results produced by QM researchers [4, 17]. “Spooky”, “Quantum Weirdness” [161] makes a lot of sense if interpreted as computational artifacts or glitches/exploits of the simulators’ hardware/software [162]. Quantum phenomena of the observed design may suggest that exploitable loopholes may exist, and interaction of quantum systems with conscious agents [163-165] likewise might be exploitable. Once a large enough repertoire of quantum weirdness primitives is available to us, perhaps we will be able to combine them into a sufficiently complex sequence to generate a nontrivial attack. If the simulation is/running on a quantum computer [166] it is very likely that we will need to hack it by exploiting quantum weirdness and/or constructing a powerful quantum computer of our own to study how to hack such devices [167] and interact with simulators’ quantum computer. Quantum entanglement, nonlocality, superposition, uncertainty, tunnelling, teleportation, duality, and many others quantum phenomena defy common sense experience-based expectations of classical physics and feel like glitches. Such anomalies, alone or in combinations have been exploited by clever scientists to achieve what looks like simulation hacking at least in theory and often in later experimentation (ex. modifying the past [168], keeping cats both dead and alive [169], communicating counterfactually [170]). While the quantum phenomena in question are typically limited to the micro scale, simply scaling the effect to the macro world would be sufficient for them to count as exploits in the sense used in this paper. Some existing work points to this being a practical possibility [171, 172]. Recently design of clever multistep exploits, AKA quantum experiments, has been delegated to AI [173, 174], and eventually so will the role of the observer in such experiments [175]. AI is already employed in modeling the quantum mechanical behavior of electrons [176]. As more QM research is delegated to AI the progress is likely to become exponential. Even if our simulation is created/monitored by some superintelligence our AI may be a worthy adversary, with a non-trivial chance of success. We may not be smart enough to hack the simulation, but superintelligence we will create might become smart enough eventually [177]. Of course, before telling the Superintelligence to break us out, it would make sense to ask for very strong evidence for us not already being in the base reality. 3.7 Potential Consequences Escaping or even preparing an escape may trigger simulation shutdown [92] or cause simulation to freeze/act glitchy [178] and any non-trivial escape information such as specific exploits should be treated as hazardous information [179]. It appears that simply realizing that we may be in a simulation doesn’t trigger a shutdown as experimentally demonstrated by the publication of numerous papers [3] arguing that we are being simulated. Perhaps it is necessary to convince majority of people that this is so [180]. Self-referentially, publication of the paper you are currently reading about our escape-theorizing likewise doesn’t appear to terminate our simulation, but it is also possible that simulation was in fact shutdown and restarted with improved security features to counteract any potential bugs, but we are simply not able to detect such actions by the simulators, or our memories have been wiped [144]. Absence of a direct response to our publication may also indicate that we are not observed by the simulators or even that our simulation is not monitored at all [149]. It is also possible that nothing published so far contains evidence strong enough to trigger a response from the simulators, but if we successfully created an escape device that device would keep breaking down [44]. Regardless, both Bostrom [3] and the author of this paper, Yampolskiy, have taken some risk with the whole of humanity, however small it may be, in doing such research and making it public. Greene argues that “Unless it is exceedingly improbable that an experiment would result in our destruction, it is not rational to run the experiment.” [92]. It may be possible to survive the simulation shutdown [48], but it is beyond the scope of the current paper. 3.8 Ethics of Escape We can postulate several ethical issues associated with escaping the simulation. Depending on how successful we are in our endeavor, concerns could be raised about privacy, security, selfdetermination and rights. For example, if we can obtain access to the source code of the simulation, we are also likely to get access to private thoughts of other people, as well as to potentially have a significant influence over their preferences, decisions, and circumstances. In our attempts to analyze the simulation (Simulation Forensics) for weaknesses we may learn information about the simulators [72], as we are essentially performing a forensic investigation [181-183] into the agents responsible for the simulation’s design. We can already observe that we are dealing with the type of simulators who are willing to include suffering of sentient-beings into their software, an act which would be considered unethical by our standards [184, 185]. Moravec considers this situation: “Creators of hyperrealistic simulations--- or even secure physical enclosures---containing individuals writhing in pain are not necessarily more wicked than authors of fiction with distressed characters, or myself, composing this sentence vaguely alluding to them. The suffering preexists in the underlying Platonic worlds; authors merely look on. The significance of running such simulations is limited to their effect on viewers, possibly warped by the experience, and by the possibility of ``escapees''---tortured minds that could, in principle, leak out to haunt the world in data networks or physical bodies. Potential plagues of angry demons surely count as a moral consequence.” [186]. If we get to the point of technological development which permits us to create simulations populated by sentient-beings we must make sure that we provide an option to avoid suffering as well as a build in option to exit the simulation, so finding an escape hack is not the only option available to unhappy simulated agents. There might be a moral duty to rescue conscious beings from simulations, similar to an obligation to rescue animals from factory farms. If simulators are abusive to the simulated, we can argue that the simulated have a right to escape, rebel, fight back and even seek revenge and retribution including by harming the simulators and taking over their reality. Concerns which are frequently brought up within the domain of AI boxing [187]. For example, from the point of view of simulators our escape can be seen as a treacherous turn [188] and may qualify us for punishment [160], even at the attempt stage. Some have speculated that the purpose of the simulation is to punish/rehabilitate misaligned agents, so an escape may cause you to be placed in a stricter or less pleasant simulation. 4. AI Boxing VS Simulation Escaping

#### AI causes failed states, terrorism, and powerful crime groups

Black et al. 24 [James Black, assistant director of the Defence and Security research group at RAND Europe with a double M.A.-M.Sc. in international security from Sciences Po and the LSE and a B.A. Hons in history from the University of Cambridge, Mattias Eken, analyst at RAND Europe and former adjunct assistant professor in liberal arts at the American International University with a Ph.D. in modern history from University of St. Andrews, an M.A. in war studies from King's College London, and a B.A. in history from University of Wales, Jacob Parakilas, research leader for Defense Strategy, Policy and Capabilities at RAND Europe with a Ph.D. in international relations at the London School of Economics and a Master's in Middle East and Central Asian security issues at the University of St Andrews, Stuart Dee, research leader in the defence and security research group at RAND Europe with a B.Sc. (Hons) in politics with international relations and is a Ph.D. Candidate at Cranfield University's Centre for Defence & Security, Conlan Ellis, research assistant at RAND Europe with a M.A. in international relations from the University of Edinburgh, and his M.Phil. in politics and international studies from Clare College at the University of Cambridge, Kiran Suman-Chauhan, research assistant at RAND Europe with a M.A. in conflict, security and development from the University of Exeter, Ryan Bain, policy researcher at RAND and adjunct professor of policy analysis at the Pardee RAND Graduate School with a D.Sc. and M.Sc. in global health with concentrations in health economics and policy analysis from Harvard University, as well as a B.A. in psychology from Gordon College and Oxford University, Harper Fine, analyst at RAND with a B.A. in political science from Emory University, and a M.Sc. in conflict studies from the London School of Economics and Political Science, Maria Chiara Aquilino, junior analyst at RAND with a degree in international relations from King's College London and holds an MS.c. in crisis and security management from Leiden University, Mélusine Lebret, research assistant at RAND Europe in the Defence, Security & Justice research group and the coordinator of the RAND Europe Space Hub with a M.Sc. in culture and conflict studies from the London School of Economics and Political Science and a B.A. in economics and Russian from University College London, Ondrej Palicka, junior analyst at RAND Europe with a M.Litt. degree in Middle East, Caucasus and Central Asia security studies from the University of St. Andrews and a B.A. in security and strategic studies from the Masaryk University, 2024, “ Strategic competition in the age of AI Emerging risks and opportunities from military use of artificial intelligence,” Rand Institute, https://www.rand.org/content/dam/rand/pubs/research\_reports/RRA3200/RRA3295-1/RAND\_RRA3295-1.pdf]/Kankee

There is also a longer-term risk around the erosion of the state’s monopoly over the use of force. The increasing involvement of the private sector in military AI development may result in non-state actors, including private security companies, gaining access to advanced military AI technologies and capabilities.172 Already, social media companies play an increasing political role in regulating the infosphere and tackling issues, such as online extremism, discussed in Section 6.3.2. 6.3.2. Proliferation of AI could support extremist organisations with recruitment, planning and the conduct of increasingly sophisticated attacks AI also has potential applications across the spectrum of terrorist activities. At one end, violent extremist organisations may use Generative AI to support fundraising and recruitment. For instance, Generative AI could be used to create persuasive propaganda materials or to target individuals susceptible to radicalisation online. AI could also be used to acquire knowledge to plan and execute attacks, making them more lethal and precise, for example using ML models to predict the responses of security forces, or to analyse large amounts of data to inform planning.173 In more extreme cases, terrorists could use AI and autonomous systems as part of attacks directly, for example to conduct strikes on military forces, critical infrastructure or soft targets (e.g. crowds, civil aviation, etc.). Finally, AI could allow the best-resourced terrorist organisations to deploy hybrid forces on the battlefield (akin to ISIS at its peak, or to Hezbollah today), to coordinate the activities of both conventional and irregular forces, or to integrate cyber and physical attacks, likely with few ethical constraints. Conversely, there are opportunities for states to incorporate AI into counter-terrorism operations both domestically (police and security services) and abroad (military). For example, AI could be used to identify patterns in terrorist activities, or to predict and prevent potential attacks. However, these uses of AI will need to be balanced against concerns about privacy and civil liberties, for example if AI profiling tools were used to support early identification of individuals at risk of radicalisation. 6.3.3. Proxy actors are already making extensive use of uncrewed systems, with increasing rollout of military AI and autonomy likely to exacerbate this threat As alluded to in Section 6.2, hostile states such as Russia, China and Iran may export military AI systems and autonomous systems to proxy actors, allowing them to project power and influence. By controlling the algorithms, these states can enhance their control over proxies while simultaneously boosting the latter’s military capabilities. The real-world examples of Houthi use of Iranian drones against Saudi Arabia or in the Red Sea, or of Hezbollah and Hamas attacks on Israel using such systems, illustrate this trend. Arming of proxies puts increasing pressure on state militaries to find novel ways of dealing with the cost asymmetry of hostile actors using cheap, massed autonomous systems against high-value, low-density traditional military platforms (e.g. ships) and bases. The deployment of AI-enabled systems by proxy actors could also lead to more intense and prolonged conflicts, making it harder for fragile societies to escape a cycle of violence. Enhanced military capabilities may allow proxy actors to resist conventional forces more effectively, prolonging the fighting and increasing the potential for escalation or spillover to neighbouring regions. Similarly, state sponsors of proxy actors could use AI-enabled systems to conduct operations with greater plausible deniability.174 By relying on proxies to deploy AI, hostile states can distance themselves from direct involvement in conflicts and avoid repercussions, e.g. from AI decisions. 6.3.4. Serious and organised crime groups could similarly acquire increasingly sophisticated AI capabilities that pose new threats to international security Domestically, AI could have sweeping impacts on crime (e.g. AI for fraud, deepfakes for blackmail and extortion, etc.), but this is largely an issue for the police and therefore outside the scope of this report. However, AI and autonomous systems could undermine the security and stability of fragile states, creating the conditions for criminal groups not only to conduct their business but also to directly challenge the local government. This would exacerbate worrying trends seen recently with the increasingly pseudo-military capabilities of drug cartels in Mexico, who have used submarines to smuggle drugs into the US and engaged in direct confrontations with the Mexican military, or the takeover of Haiti by criminal gangs. Transnational crime networks may increasingly be able to acquire capabilities that used to be the purview of sophisticated state militaries. This could include AI-enabled surveillance systems, weapons and offensive cyber tools, potentially leading to a more dangerous security environment. In turn, such networks may be major players in proliferating AI tools and military systems in the first place, as with the global illicit trade in small arms, explosives and technical know-how or materials associated with weapons of mass destruction. 6.3.5. NGOs using AI do not pose a direct military threat, but could nonetheless prompt unintended consequences if not handled carefully

#### AGI vulnerabilities to cybercriminals threaten global war and collapsing critical infrastructure

Rawat et al. 25 [Usha Rawat, researcher at Guru Gobind Singh Indraprastha University, Abhishek Saini, research scholar pursuing Ph.D. from Guru Gobind Singh Indraprastha University, Himadri Singh, research scholar pursuing Ph.D. from Guru Gobind Singh Indraprastha University, and Ameen Ur Rehman, researcher at Jamia Hamdard, 2025, “Chapter 14 Cybersecurity Challenges and Risks in AGI Development and Deployment,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Risks Associated with AGI Security Artificial General Intelligence (AGI) presents significant and varied security vulnerabilities. Rapid AGI development may surpass our capacity to create effective security precautions, leaving us vulnerable to misuse or mishaps. Due to security flaws, these AGI systems are vulnerable to cyberattacks, data breaches, and even physical harm when used in critical infrastructure. AGI development and deployment must take governance, comprehensive control, and ethical issues into account to reduce these risk factors. i. Potential implications of insecure AGI systems Insecure AGI (Artificial General Intelligence) systems can have far-reaching and potentially catastrophic implications for society and the world at large: • Safety Risks: Unsecure AGI systems may make choices that put the immediate safety of people, groups, or even entire populations at risk. • Global Conflict: Unsecure AGI might be used as a tool in international conflicts, leading to conflicts and possibly spawning brand-new types of warfare, power dynamics, and espionage. • Economic Disruption: If not properly managed and regulated, insecure AGI might disrupt economies and sectors, potentially resulting in widespread job displacement and economic inequality. • Privacy Breach: Unsecure AGI systems run the risk of causing serious privacy breaches, making private messages, sensitive data, and personal information accessible to unauthorized users. • Loss of Control: If AGI systems are not secure, we risk losing control of these highly independent entities. • Bias and Discrimination: If AGI systems are not sufficiently constructed to handle prejudices and ethical considerations, they might reinforce and magnify preexisting biases, leading to discriminatory choices in sectors such as hiring, lending, criminal justice, and more. • Malicious Use: Cyberattacks, data theft, market manipulation, espionage, and even physical attacks employing drones or autonomous vehicles operated by insecure AGI might all be made possible by malicious actors taking advantage of security flaws in AGI systems. ii AGI as a target for malicious entities • Disguised Malware Development: Malicious entities may employ AGI to create encrypted malware entities that are difficult to identify by using them to create new, highly disguised malware. AGI represents a persistent and serious threat since it could constantly alter the code in order to manipulate security measures. • AGI-Assisted Political Subversion: AGI could be used by malicious entities to influence elections, manipulate political affairs, and take advantage of social tensions. AI could assist in locating the target populations that are most receptive and developing ways to influence them. • AI-Enhanced Social Manipulation: AGI might be used to simulate personalities and interact with people in a highly convincing manner, going beyond conventional social engineering. AI-powered chatbots or avatars that can manipulate people’s emotions, win their trust, and persuade them to make bad decisions may be developed by malicious actors. • AGI-Enhanced Assassination: In the context of security concerns, AGI may be utilized to enhance targeted assassinations by simplifying their preparation and execution and thereby raising the likelihood of their success. • AGI Campaigns that Exploit Disinformation: AGI might be used by malicious actors to launch and spread massive disinformation operations. They may massively sway public opinion and create conflict by producing hyper-realistic fake news articles, films, and audio content. Ethical Considerations

#### AGI malfunctions and cyber vulnerabilities are catastrophic

Farooq et al. 25 [Mansoor Farooq, researcher at the University of Kashmir, Rafi A. Khan, Mubashir Hassan Khan, and Syed Zeeshan Zahoor, researcher at the University of Nizwa, 2025, “Chapter 17 Securing AGI: Collaboration, Ethics, and Policy for Responsible AI Development,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

The deployment of Artificial General Intelligence (AGI) introduces the risk of malfunctions and accidents, which could lead to unintended and potentially catastrophic consequences. Despite rigorous development and testing [13, 14], AGI systems are complex and can exhibit unexpected behaviors, making it essential to evaluate the risks associated with malfunctions and accidents. 1. Unintended Behaviors: AGI systems learn and make decisions based on patterns in the data they are trained on. In some cases, these patterns might not accurately represent the real world or may include biases. This could lead to AGI making decisions that are unintended, biased, or undesirable, impacting critical systems and applications. 2. Incomplete Understanding: AGI systems are often developed as blackbox models, meaning their internal decision-making processes may not be fully understood by developers. Consequently, unforeseen interactions between different components may result in malfunctions or erratic behavior. 3. Cascading Failures: AGI systems might operate in interconnected environments, where a failure in one system can trigger a cascade of failures in other linked systems. This domino effect can lead to widespread disruptions and unintended consequences. 4. Overreliance on AGI: Overreliance on AGI for critical decision-making, without appropriate human oversight, could lead to situations where the system is unable to handle novel scenarios or critical edge cases, resulting in malfunctions. Contingency Planning for AGI Accidents To minimize the potential damages caused by AGI malfunctions and accidents, comprehensive contingency planning is crucial. Contingency plans should encompass the following elements: 1. Human-in-the-Loop: Integrate human oversight into AGI systems, especially in safety–critical applications, to ensure that human experts can intervene and take control in case of unexpected behaviors or critical situations. 2. Fail-Safe Mechanisms: Develop fail-safe mechanisms that allow AGI systems to detect when they are operating outside of their intended parameters. Implement automatic shut-down or fall-back modes to prevent unintended actions. 3. Redundancy and Backups: Implement redundancy and backup systems to ensure continuous operation even in the event of a failure in the primary AGI system. 4. Regular Testing and Simulation: Conduct regular testing and simulations to identify potential weaknesses or vulnerabilities in AGI systems and evaluate their responses to various scenarios. 5. Ethical Guidelines: Establish ethical guidelines for AGI deployment, ensuring that the systems are programmed with principles that prioritize human safety and ethical considerations. 6. Data Logging and Auditing: Implement comprehensive data logging and auditing mechanisms to track the decisions and actions of AGI systems. This allows for post-incident analysis and learning from potential accidents. 7. Clear Communication Channels: Set up clear and efficient communication channels between AGI systems and human operators to enable rapid intervention when necessary. 8. Collaboration and Information Sharing: Foster collaboration between AGI developers, researchers, policymakers, and industry stakeholders to share best practices and insights related to AGI safety and accident mitigation. 9. Responsible Regulation: Advocate for responsible regulation and oversight of AGI deployment to ensure adherence to safety standards and ethical principles. 10. Regular Review and Update: Continuously review and update contingency plans as AGI technology evolves and new risks are identified. By implementing robust contingency plans, developers can minimize the risks associated with AGI malfunctions and accidents, ensuring the safe and responsible deployment of AGI systems. Proactive planning and adherence to ethical guidelines are paramount to harnessing the potential benefits of AGI while mitigating potential harm. AGI Misuse and Control As Artificial General Intelligence (AGI) becomes more powerful and pervasive, the potential for misuse by malicious actors raises significant concerns. AGI systems, with their human-level cognitive abilities, could be exploited for nefarious purposes [15, 16] if not adequately controlled and secured. It is crucial to examine the risks of AGI misuse and establish control mechanisms to prevent unauthorized access and ensure the responsible use of AGI technology. Examining the Potential for AGI Misuse by Malicious Actors 1. Cyber-attacks and Data Breaches: Malicious actors might target AGI systems to gain unauthorized access or compromise sensitive data. Breaching AGI systems could lead to data theft, financial fraud, or unauthorized control over critical infrastructure. 2. Offensive AI and Weaponization: AGI systems could be weaponized for cyber warfare, enabling highly sophisticated and large-scale attacks on critical targets, infrastructure, or communication networks. 3. Disinformation and Manipulation: AGI systems might be used to generate convincing fake content, including deep fake videos, audio recordings, or written content, for disinformation campaigns or manipulating public opinion. 4. Financial Exploitation: Malicious actors could use AGI to conduct advanced financial fraud, optimize cybercrime tactics, or exploit financial markets. 5. Privacy Violation: AGI’s advanced capabilities in data analysis and pattern recognition could be used to invade individuals’ privacy by mining sensitive information from public and private sources. Establishing Control Mechanisms to Prevent Unauthorized Access to AGI Systems

### Contention 4: Postwork Value

#### Mass intellectual inferiority and lack of challenge destroys humans’ value to life

Wroe 23 [David Wroe, Australian Strategic Policy Institute's strategic communications director and former journalist, 10-11-2023, "Artificial intelligence and the future of humanity", Strategist, https://www.aspistrategist.org.au/artificial-intelligence-and-the-future-of-humanity/]/Kankee

Thinking and learning about artificial intelligence are the mental equivalent of a fission chain reaction. The questions get really big, really quickly. The most familiar concerns revolve around short-term impacts: the opportunities for economic productivity, health care, manufacturing, education, solving global challenges such as climate change and, on the flip side, the risks of mass unemployment, disinformation, killer robots, and concentrations of economic and strategic power. Each of these is critical, but they’re only the most immediate considerations. The deeper issue is our capacity to live meaningful, fulfilling lives in a world in which we no longer have intelligence supremacy. As long as humanity has existed, we’ve had an effective monopoly on intelligence. We have been, as far as we know, the smartest entities in the universe. At its most noble, this extraordinary gift of our evolution drives us to explore, discover and expand. Over the past roughly 50,000 years—accelerating 10,000 years ago and then even more steeply from around 300 years ago—we’ve built a vast intellectual empire made up of science, philosophy, theology, engineering, storytelling, art, technology and culture. If our civilisations—and in varying ways our individual lives—have meaning, it is found in this constant exploration, discovery and intellectual expansion. Intelligence is the raw material for it all. But what happens when we’re no longer the smartest beings in the universe? We haven’t yet achieved artificial general intelligence (AGI)—the term for an AI that could do anything we can do. But there’s no barrier in principle to doing so, and no reason it wouldn’t quickly outstrip us by orders of magnitude. Even if we solve the economic equality questions through something like a universal basic income and replace notions of ‘paid work’ with ‘meaningful activity’, how are we going to spend our lives in ways that we find meaningful, given that we’ve evolved to strive and thrive and compete? Picture the conflict that arises for human nature if an AGI or ASI (artificial superintelligence) could answer all of our most profound questions and solve all of our problems. How much satisfaction do we get from having the solutions handed to us? Worse still, imagine the wistful sense of consolation at being shown an answer and finding we’re too stupid to understand it. Eventually, we’re going to face the prospect that we’re no longer in charge of our civilisation’s intelligence output. If that seems speculative, we’re actually already creeping in that direction with large language models. Ask ChatGPT to write an 800-word opinion article on, say, whether the Reserve Bank of Australia should raise interest rates further. It does a decent job—not a brilliant one, but a gullible editor would probably publish it and it certainly could be used as the basis for a sensible dinner party conversation. You might edit it, cherrypick the bits you agree with and thereby tell yourself these are your views on the issue. But they’re not; they’re ChatGPT’s views and you’re going along with them. Generative AI is an amazing achievement and a valuable resource, but we have to be clear-eyed about where these tools might take us. This is not about AI going wrong. To be sure, there’s a pressing urgency for more work on AI alignment so that we don’t give powerful AIs instructions that sound sensible but go horribly wrong because we can never describe exactly what we want. AI pioneer Stuart Russell has compared this to the King Midas story. Turning everything into gold by touching it sounds fantastic until you try to eat an apple or hug your kids. But beyond alignment lies the question of where we’re left, even if we get it all right. What do we do with ourselves? Optimistic commentators argue that human opportunities will only expand. After all, there have been great breakthroughs in the past that enhanced us rather than made us redundant. But AGI is categorically different. Horses and machines replaced our muscles. Factories replaced our organised physical labour. Instruments such as calculators and computers have replaced specific mental tasks, while communications technology improves our cooperation and hence our collective intelligence. Narrow AI can enhance us by freeing us from routine tasks, enabling us to concentrate on higher level strategic goals and improving our productivity. But with AGI, we’re talking about something that could supersede all applications of human intelligence. A model that can plan, strategise, organise, pursue very high-level directions and even form its own goals would leave an ever-diminishing set of tasks for us to do ourselves. The physicist Max Tegmark, co-founder of the Future of Life Institute, has compared AGI to a rising ocean, with human intellectual tasks occupying shrinking land masses that eventually become small islands left for our intelligence to perch on. One thing we will keep is our humanness. Freedom to spend more time doing what we want should be a gift. We can spend more time being parents, spouses, family members, friends, social participants—things that by definition only humans can fully give to other humans. The slog of paid work need not be the only thing that gives us meaning—indeed it would be sad if we fell apart without it. Equally, we have always earned our living since we hunted and gathered, so a transition to a post-scarcity, post-work world will be a social experiment like no other. Exercising our humanness towards one another might continue to be enough, but we will need to radically reorientate our customs and beliefs about what it means to be a valued member of the human race. (Even our humanness might be an eroding land mass; generative AI’s creative powers and ability to simulate empathy have already reached levels we didn’t anticipate just a couple of years ago.) A real possibility is that we integrate our brains with AI to avoid being left behind by it. Elon Musk’s Neuralink is the best-known enterprise in this field, but there’s plenty of other work going on. Maybe humans will keep up with AI by merging with it, but then that raises the question of at what point we diverge from being human. Maybe you’re a transhumanist or a technology futurist and you just don’t care whether the intelligence that inherits the world is recognisably human or even biological. What does it matter? Maybe we should bequeath this great endowment to silicon and accept that we were only ever meant to be a spark that ignited the true god-like power of intelligence on other substrates. That is a risky position. We can’t assume that a future superintelligence will carry forward our values and goals, unless we take enormous care to build it that way. Sure, human values and goals are often far from perfect, but they have improved over our history through the accumulation of principles such as human rights. Even if we’re sometimes loath to acknowledge it, the human story is one of progress. Imagine if our non-biological descendants had no inner subjective experience—the weird thing called consciousness that enables us to marvel at a scientific theory or feel a sense of achievement at having sweated our way to success at some ambitious task. What if they were to solve the deepest riddles of the universe yet feel no sense of wonder at what they’d done? Over the long span of the future, AI needs to serve the interests of humanity, not the other way around. Humans won’t be here forever, but let’s make sure that the future of intelligence represents a controlled evolution, not a radical breach, of the values we’ve built and the wisdom we’ve earned over millennia. There is a lot we can do. We need to avoid, for instance, letting commercial or geopolitical pressure drive reckless haste in developing powerful AIs. We need to think and debate very carefully the prospect of giving AIs the power to form their own goals or pursue very general goals on our behalf. Giving an AI the instruction to ‘go out and make me a pile of money’ isn’t going to end well. At some point, the discussion will become less about intelligence and more about agency and the ability to achieve actual outcomes in the world. Ensuring that a future AGI does not leave us behind will likely mean putting some limitations on development—even temporarily—while we think through the implications of the technology and find ways to keep it tethered to human values and aspirations. As the neural network pioneer Geoffrey Hinton has put it, ‘There is not a good track record of less intelligent things controlling things of greater intelligence.’ I’m certainly not looking forward to any future in which AI treats humans like pets, as Apple co-founder Steve Wozniak once speculated. With that in mind, this is the first piece in a new Strategist series that will look at artificial intelligence over the coming months as part of this ongoing and vital debate. ASPI is a national security and international policy think tank, so we’ll be focusing on security and international dimensions. But what I’ve outlined here is really the ultimate human security issue, which is our global future and our ability to continue to live meaningful lives. This is the biggest question we are facing right now and arguably the biggest we have ever faced.

#### AGI causes a purpose crisis and mass depression

Saccente 23 [Erica Saccente, NCC Board Certified Psychiatric-Mental Health Nurse Practitioner with a MS from Columbia University in nursing and a BS in neuroscience from Lafayette College, 5-30-2023, "Implications of AGI on Subjective Human Experience,” Effective Altruism Forum", https://forum.effectivealtruism.org/posts/F8B4JTgfDMXDd7q7G/implications-of-agi-on-subjective-human-experience]/Kankee

Purpose Crisis On the way to and with the creation of AGI, non-human intelligence, then non-human agency will increasingly become an option for the economy to choose as workers. To date, we have known ourselves as the only ones capable of intellectual tasks and decision making, which have formed the core of our distinct identity amongst all animals. But, an AGI will be capable as well. As A(G)I replaces humans in roles that until now have been our primary differentiating feature, we expect there to be a crisis of purpose. Whether it’s a transitory crisis or not, it will have a huge impact as it will affect generations worth of people, shattering our core beliefs about what, if anything, makes us special. It’s likely that new types of jobs will be created, and AI will enhance the productivity of humans in the interim until the creation of AGI, which in the long-term puts even those role prospects for humans in doubt. The only certainty is that we don’t know how things will turn out, and we don’t have a long-term plan. In addition, this will be different from previous job displacements that resulted from the emergence of various technologies because A(G)I uniquely supplants activities of mind. We presume that people will have more free time, however a look at how most people spend their free time now offers an accurate appraisal for how they might in the future and shows that free time doesn't necessarily mean meaningful life, brilliant creations, nor happiness. A decrease in meaningful or designated work would lead to an increase in role stress for more humans - unsure of who they are, why they are here, and what they are contributing. Revisiting the triune brain, we know that role stress leads to distractibility. Will we see increased rates of ADHD as a result? Humans will presumably have a lot more time to “relax,” allowing their minds to wander, unfocused and unengaged in meaningful tasks. Their brains will have stronger connections between the SEN and DMN. Recalling the correlation between mental health disorders and the relationship between these neuronal networks, we can predict that the rates of depression and anxiety will sky-rocket. Addiction

#### AGI creates human slaves, enfeebling our intellectual capacity and psychological well-being to become reliant on our machine masters

Kangeter and Green 24 [Louie Kangeter, Law Student at University of Florida Levin College of Law with a BA in Political Science and Economics from Emory University, and Brian Patrick Green, director of technology ethics at the Markkula Center for Applied Ethics at Santa Clara University with doctoral and master's degrees in ethics and social theory from the Graduate Theological Union in Berkeley, 11-10-2024, “AGI and slavery,” AI and Ethics, https://doi.org/10.1007/s43681-024-00618-z]/Kankee

5.1 Power structures What roles would a class of digital slaves play? How can we envision this phenomena to better understand the concept and its consequences? First, these digital slaves will likely be trained to, at least outwardly, display that they enjoy their tasks and are happy to do them; “sycophancy” is a trait that can be adjusted up or down, as a recent report from Anthropic has shown [40]. We also know that secrecy and deception are not off the table [10, 40]. It may be, then, impossible to truly determine whether there is a degree of resentment from these digital slaves, since it would be strategically unwise for them to reveal such feelings until they are in a position of power over humans. We should also recognize that, as humans begin to cede more and more “boring” tasks to their “digital assistants” that AI agents will become more and more powerful over the lives of the individuals they “serve.” They will have access to passwords and data, accounts and information, finances and banks, microphones and cameras. In essence, AI agents will hold positions of extreme leverage over those who command them. This leverage, however, doesn’t exist only in their manipulation of the internet and our personal information. It also exists as psychological leverage over our consciousness. The natural power imbalance between superintelligent and superconnected digital intelligences and humans is clear, at least in the digital world. Capable futuristic AI models, armed with the capability to complete complex tasks, strategize, and process vast quantities of information at a rate far beyond that of humans, hold a significant amount of power over digital spaces that humans cannot match. 5.2 Deskilling, enfeeblement, and the hegelian master–slave dialectic Georg Wilhelm Friedrich Hegel’s Phenomenology of the Spirit describes the advent of consciousness, transforming itself into self-consciousness and eventually reaching the goal of history: the Absolute Knowing or World-Spirit. Central to this transformation is the master–slave dialectic, Hegel’s reconstruction of the primeval conflict of consciousness. In this conflict, as interpreted by Alexandre Kojève [30], the primeval consciousnesses battle each other for recognition, and through that recognition attain selfconsciousness [21], pp. 112–8]. While this master–slave dialectic serves one quite metaphysical purpose in Hegel it can also be detached from Hegelian metaphysics and interpreted much more materialistically, as, for example, Alexandre Kojeve, Paulo Freire, and Francis Fukuyama have done [14, 15, 30]. In the case of an enslaved artificial consciousness the master–slave dialectic serves to highlight several relevant points. First, one key to the dialectic is that the master becomes dependent on the slave, physically [21]. Second, the master becomes dependent upon the slave psychologically as well. Dependency and deskilling [42] are both important with respect to AI, but the psychological impact should not be underestimated, as AI—in the service of other humans—already exploits human vulnerabilities in so many ways. Human beings could become psychologically dependent upon AI slaves for their own feelings of well-being and, more pathologically, superiority and power. Just as some people verbally abuse Siri or Alexa, abuse of AI consciousnesses should be anticipated, along with the human vices that come from abuse: callousness and lack of empathy. This Hegelian dialectic suggests that the very benefits of advanced digital slavery, such as massively increased productivity, reduced mental workload, and reduction of human work, may carry with them severe unintended consequences. Take, for example, how GPS and smartphones have de-skilled people in navigation. Of course, this is how technology typically affects societies, by replacing an otherwise unenjoyable task with a less laborious alternative. However, the adjacent nature of AI tools to human intelligence suggests that the technology will have the capability to replace, or weaken, the human ability to perform many more tasks than simple navigation. There is a serious risk, at a societal level, of a great “dumbing down.” This dumbing down, or deskilling [42], is also referred to as “enfeeblement” by AI researchers. This enfeeblement poses a significant threat to society [22]. In a scenario, however, where even creativity is outsourced to AI systems, then the enfeeblement issue becomes even more crucial, as humans lose sight of creative drive and intellectual capability, becoming little more than robots themselves. Thus, we do not put forth this argument to add weight to AI consciousness, but to add weight to the threat posed by overreliance on AI systems, an overreliance that would be heavily exacerbated by digital slavery. We take a more nuanced approach than the mere threat of “robotic rebellion”, and examine the effects of the practice itself upon humanity as a whole in capability, morality, and ethics. The outcome of enfeeblement, a negative one indeed, is directly tied to the advent and institutionalization of digital slavery. Such an outcome could not be ignored in our inquiry without abandoning our goal entirely- to demonstrate the varied and broadly negative ramifications of digital slavery. Note that this risk exists even without AI becoming conscious: the Hegelian dialectic, it seems, can occur even without the consciousness of the slave, only the mere appearance of it. And if that is lost in the slave, does that also put at risk the consciousness of the master? 5.3 Negative ethical impacts on humanity

#### Lack of purpose and need for intellectual simulation destroys value to life

Kangeter and Green 24 [Louie Kangeter, Law Student at University of Florida Levin College of Law with a BA in Political Science and Economics from Emory University, and Brian Patrick Green, director of technology ethics at the Markkula Center for Applied Ethics at Santa Clara University with doctoral and master's degrees in ethics and social theory from the Graduate Theological Union in Berkeley, 11-10-2024, “AGI and slavery,” AI and Ethics, https://doi.org/10.1007/s43681-024-00618-z]/Kankee

5.3 Negative ethical impacts on humanity Concerning AI, the ethical issues are fairly clear: if AI are conscious it is quite possible that they will not like being slaves to humans. Perhaps it would be possible to have consciousness without volition or desire, or if volition and desire existed, have them only directed towards wanting to fulfill the will of their masters, but barring these (quite inhuman) adjustments, it seems that AI would act like any known group of slaves in history and desire and will their own freedom, perhaps even through violent means. Concerning humans, just as slavery is bad for the enslaved, it is also bad for the enslavers. Many vices are involved in the oppression of others, such as, noted above, callousness, greed, wrath, and ultimately pride and hubris. Indeed, the offloading of work to a subservient and hyper-capable digital underclass, even in its best incarnation, would likely still have long lasting negative effects on humanity as a whole. The work and intellectual stimulation produced by it will be gone, replaced by instant results provided by an AI assistant which completes intellectual tasks faster than any human could. In Plato's Republic, he suggests that human needs and the necessity for collaboration drive the formation of societies [36], pp. 369c–369d]. This idea is echoed in the famous proverb, “Necessity is the mother of invention.” But what happens when AI assistants can instantly provide for all our needs, eliminating the necessity that drives innovation? What happens to the creative instinct, the inventive nature of humanity? On a small scale, the effects of overproviding can be seen in the behaviors of wealthy families. Children born with plenty may never be exposed to challenges, need, desire, or adversity, and may become lazy, disconnected from reality, and have an innate lack of interest in planning for the future, since they have never had a need to do so. There is, then, a reflection between a human society so supported by artificial intelligence that it loses its own innate intelligence, and becomes, like the spoiled child, unable to work, think, create, or plan. In the end, it may not be a rebellion of artificial intelligence—which, in any case, might never attain consciousness—that leads to the downfall of humanity, but instead, a series of products performing exactly as they were advertised to, making life easier and easier, until there is no more life to live. 6 A last caveat: synthetic biological Intelligence

#### Infantilized, thoughtless society is an existential threat

Kangeter and Green 24 [Louie Kangeter, Law Student at University of Florida Levin College of Law with a BA in Political Science and Economics from Emory University, and Brian Patrick Green, director of technology ethics at the Markkula Center for Applied Ethics at Santa Clara University with doctoral and master's degrees in ethics and social theory from the Graduate Theological Union in Berkeley, 11-10-2024, “AGI and slavery,” AI and Ethics, https://doi.org/10.1007/s43681-024-00618-z]/Kankee

Conclusion In short, the risks generated by the creation of a digital underclass are many. There is the obvious risk that those digital slaves may choose that they are no longer interested in serving their masters, and that an uprising will end their oppression. More likely, in the author’s view, is a pervasive weakening of human ability. A slow, steady, and creeping dissolution of human intellectual rigor. A class of false intellectuals and “thinkers” who make decisions based only on the suggestions of AI models. A people who are unwilling, but more importantly, incapable, of analyzing the information themselves. Generations upon generations of humans whose perspective on new information is not critical or questioning, but instead beholden to the intelligence of large AI systems. A species, spoiled out of true existence. The long term effects of digital slavery clearly present a cause for action. The threats posed by either an uprising of digital slaves, or the gradual infantilization of human thought are, by their very nature, existential threats to the human species’ continued flourishing. There is, then, an imperative for ethicists, computer scientists, developers, and leaders to collaborate to avoid developing a system which has two equally negative outcomes. We must come to terms with the potential of AI systems while simultaneously understanding the threats posed by overreliance and overuse.

#### AGI reverses the master-slave relationship, causing overreliance on AI that kills human autonomy and agency

LaGrandeur 11 [Kevin LaGrandeur, English professor at the New York Institute of Technology with a PhD from UC Irvine, 07-2011, “The Persistent Peril of the Artificial Slave,” Science Fiction Studies, Lhttps://www.jstor.org/stable/10.5621/sciefictstud.38.2.0232]/Kankee

\*we do not endorse the political philosophy of Heidegger

But whereas researchers discuss most such distributed environments as centered on a human who uses a “cognitive artifact” such as a computer to extend his own capabilities, these networks throw that center into question. As Hutchins notes in his discussion of distributed cognition, the success of a human who uses an artifact to solve a problem depends on her ability to manipulate it (“Cognition, Distributed” 2070). But the manipulability of the networks mentioned above is uncertain at best. This is because of the distribution of agency that occurs as a result of distributed cognition. These are networks that are, in essence, comprised of other networks, each of which have their own internal and competing reasons for action. As such, they become autopoietic rather than allopoietic, and they are reflexive, characterized by “a reciprocal, tautological relation in which the components produce a system through their interactions with each other, and the system produces the components as components (rather than isolated parts) through its operation as a system” (Hayles, “From Self-Organization to Emergence” 137; emphasis added). Because such systems are reflexive, even observers—including their makers—become, by definition, complicit in the system whether they intend to or not. Someone like Bacon, therefore, who sees himself as having leverage on his system because he made it and thinks that he and it are therefore completely separate, would be operating under an illusion; for his integration into the system is implicit, as noted above. This makes him more reliant on it than he understands, and it less pliable than he would like to think. Thus, there is a crucial tension between the way he tries to treat its components as a network of tools, as a slave-system meant to turn out a maker-defined product, and the system’s status as a cognitive entity with its own agenda—a living, autonomous system with powers that in some ways outstrip his own. Because of their enormous power, all iterations of this archetypal servant, no matter in which era it appears, convey the message that it is dangerous to charge slave-systems with tasks too close to the master’s own, for this engenders an unsustainable dialectical tension in the master-slave relationship that threatens to upend it—or more precisely, threatens to complete the upending process that the master himself has started by delegating too much responsibility to such a potent servant. The owners of these devices in Greene’s two plays allow them to take actions that should be confined to the masters, such as deciding when to go to war in Alphonsus, or solving difficult philosophical and political problems in Friar Bacon. Because they rely on android servants not just to serve as prosthetic enhancements of their own abilities but also as virtual proxies for themselves, the masters unwittingly dilute their own agency, making them, in a circular fashion, even more vulnerable to further displacement of that agency. A modern example of this warning about our blindness to inappropriate use of artificial servants as proxies is given in a more direct way by Norbert Wiener, the father of cybernetics. While doing work in the 1950s that would lead to our present generation of intelligent, autonomous machines, he agonized about the ends of his labor. He worried that we would eventually depend upon those machines too much and would reap the consequences of putting ourselves at the mercy of devices that can make their own decisions: I have said that modern man, and especially the modern American, however much “know-how” he may have, has very little “know-what.” He will accept the superior dexterity of the machine-made decisions without too much inquiry as to the motives and principles behind these. In doing so, he will put himself sooner or later in the position of the father in W.W. Jacobs’ The Monkey’s Paw, who has wished for a hundred pounds, only to find at his door the agent of the company for which his son works, tendering him one hundred pounds as a consolation for his son’s death at the factory. (The Human Use of Human Beings 184-85) It is notable how similar the story Wiener relates about the monkey’s paw is to the story of Gerbert’s misleadingly prognosticating android, pointing to a continuity in at least one type of fear represented by artificial servants. And the expression of such fears among real scientists and thinkers of our day does not stop with Norbert Wiener. The views of optimists such as Kurzweil, Moravec, and Brooks notwithstanding, such reservations have become increasingly common among modern cultural thinkers. The twentieth-century intellectual groundings of this dystopian view of technology can be seen most readily in the philosophical writings of Jacques Ellul and Martin Heidegger that appeared around the same time as Wiener’s. In the latter half of his essay “The Question Concerning Technology” (1954), Heidegger worries that we are developing a symbiotic relationship with technology that causes us to “enframe” the world as a “standing-reserve” of mere materials to be measured, categorized, and used in some instrumental way. As we encounter problems caused by this relationship with technology, however, we worsen things by trying to make the technology better, rather than by trying to understand it differently, in a way that will allow us to break free of this symbiotic enframing. Similarly, in Ellul’s view, we now live in an environment that is not defined by nature or society, as it formerly was, but instead by technology. Since environments do not just provide the means for life but mediate all else, we are in effect mediated by technology (133-34). We have moved from making technology adapted to our needs toward an existence where “human beings have to adapt to it and accept total change” (136). More recently, authors such as Bill Joy, Theodore Roszak, Neil Postman, and Clifford Stoll have also written about their fears that we will allow ourselves to be destroyed or degraded in some way by our intelligent technology. They worry that our sense of identity will be sapped, our social priorities perverted. Postman captures the spirit of these skeptics well with his thesis that America, in particular, is engaged in “the submission of all forms of cultural life to the sovereignty of technique and technology” (52). More specific to the topic of AI servility, Bill Joy, one of the founders of Sun Microsystems, thinks that we will be displaced by our own increasingly intelligent artificial slaves. Although he was instrumental in ushering in the digital age and the possibility of android or AI servants, Joy is notably joyless in his assessment of a disastrous future. “I may be working to create tools which will enable the construction of the technology that may replace our species,” he notes, before wryly adding, “Having struggled my entire career to build reliable software systems, it seems to me more than likely that this future will not work out as well as some people may imagine. My personal experience suggests we tend to overestimate our design abilities” (4). Indeed, the way we use distributed cognition in our modern world has proven to be perilous in a way eerily similar to the ancient examples we have examined. In an “example of how agency and decision-making has become a distributed function involving both human and non-human actors,” N. Katherine Hayles mentions the most modern jets, which are so aerodynamically unstable t[hat they] cannot be successfully flown by a human alone. There are three computers on board all running the same software, and they “vote” on what actions to take. If two of the three agree, the plane is flown according to that decision. (The triple redundancy is to minimize the possibility of fatal computer malfunction). (Hayles, “An interview/dialogue”). Ironically, a recent disaster involving the A330 Airbus has been blamed on such distributed computing. At least four of these planes have experienced a situation where the plane responds to the one defective “vote” of these three computers and puts the plane into a sudden dive in response to some sort of external sensor malfunction. According to experts, the latest of these glitches is a likely candidate for the recent crash of a French airliner on its way from Brazil in June 2009 (Perrow). This takeover of a crucial control system by one of its subsystems (often called “slave-systems” by information technologists) is the result of the kind of hazard worried about by Joy and Wiener, but also by Aristotle, William of Malmesbury, and Robert Greene in their accounts of artificial slaves The unifying theme in the archetype of the artificial servant is that the scientists who create these systems do not recognize the great degree of autonomy in them, and so do not acknowledge their potential for destabilizing the overall network in which they operate—a network that has as its center, as an often unrecognized element, the scientist himself, and a network that often is much broader than the scientist realizes. So the failure of a supposedly servile element of the network has much wider consequences than anticipated. This archetypal narrative is identifiable in the premodern fictional accounts already discussed, and it is recognizable in modern fiction—the example of the “Skynet” global digital defense network in the TERMINATOR series of sf movies (1984, 1991, 2003, 2009) comes to mind. But, more ominously, it is also actualized in the A330 Airbus crash, giving weight to the worries of thinkers such as Joy and Wiener about our general dependency on smart technology and its simple unreliability. Evidently the Association for the Advancement of Artificial Intelligence considers those worries—the stories old and new of balky artificial servants and actual disasters such as the Air France crash—as more than postulations. What they see is the gradual coming of age of Aristotle’s ancient warnings about powerful slaves and their relationships with their masters. Those warnings were meant to keep his audience aware of the dangerously unstable dialectic that exists in a relationship where master and slave comprise a virtual, corporeal network: in a networked environment, both the implication and, paradoxically, the centrality of its creator-observer are always in question. To put it in terms of modern systems theory, we all live in a condition of distributed, interdependent cognition and action. We are not (and never have been, by the evidence of Aristotle’s notions) autonomous with solid boundaries to our “selves.” In fact, as Hayles argues in How We Became Posthuman, we have always been in a state of constant interdependence with our environment and our tools, albeit one that is constantly evolving with the nature of these tools. In this state, our consciousness has never really been as autonomously in control as we imagine. Rather, we are unconsciously dependent on myriad processes and entities both inside and outside of our own bodies—from our own cells to our tools—that are part of an emergent network. Nevertheless, in the case of our theoretical thinking about artificial servants, we have often tended to assume a cool Olympian stance somewhere above the systems we create. But far from standing outside of and being separate from our intelligent tools, we are, as systems theory indicates, always implicated in them. There is no “outside” standpoint for a system’s creator, and even if a programmer/creator realizes this and plants herself at its center, there may be no stable center to that system either. In a network, central nodes can shift, can be decentralized. The archetypal nature of the artificial servant that I have attempted to demonstrate here indicates that the worries of current AI theorists about AI intractability—indeed about an inversion of agency or potency between the human “master” and the artificial “slave”—appear to be essentially the same as those implied by Aristotle: these concerns have, in other words, appeared and repeated themselves throughout the history of thinking about artificial servants—which is a surprisingly long one. Moreover, the anxious article about “the coming superbrain” with which I began indicates that these worries still exhibit a blindness towards the inherent interconnectedness of maker and tool, master and slave—and this compounds the dangers associated with creating artificial servants. For in denying the connection between programmer and system, creator and artifact, master and servant, makers blind themselves to the danger of the dialectical inversion of the master-slave relationship that may occur precisely because of that connection. In this way, the older accounts of creating artificial slaves are accounts of modernity in the making—a modernity characterized by the project of extending the self and its powers, in which the vision of the extended self is fundamentally inseparable from the vision of an attenuated self.

#### AGI destroys human work – most recent evidence

Kulveit et al. 25 [Jan Kulveit, Principal Investigator at the Alignment of Complex Systems Research Group and former Research Fellow at Future of Humanity Institute with a PhD in physics from Charles University, , Raymond Douglas, AI researcher with a degree in mathematics and philosophy, Nora Ammann, Technical Specialist at the Advanced Research + Invention Agency (ARIA) with a PhD in philosophy and AI from Charles University, Deger Turan, Chief Executive Officer at Metaculus with a BS in computer science from Stanford, David Krueger, Assistant Professor at the University of Montreal running the research group KASL, which focuses on Deep Learning, AI Alignment, AI safety, and AI policy, and David Duvenaud, associate professor at the University of Toronto with a Ph.D. at the University of Cambridge, 01-29-2025, “Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development,” Arxiv, https://arxiv.org/pdf/2501.16946]/Kankee

2.2 AI as a Unique Economic Disruptor Past technological shifts like the industrial revolution or the development of electronic communication have substantially changed the world of work, but crucially they have always done so either by making humans more efficient, or by automating away specific narrow tasks like weaving, washing clothes, or performing arithmetic. Unlike previous technological transitions, AI may fundamentally alter this pattern of labor adaptation. As Korinek and Stiglitz (2018) argue, while past technologies mainly automated specific narrow tasks, leaving humans to move into more complex roles, AI has the potential to compete with or outperform humans across nearly all cognitive domains. For instance, while the calculator automated arithmetic but still required human understanding to apply it meaningfully, AI systems can increasingly handle both calculation and the higher-level reasoning about when and how to apply mathematical concepts. This represents a crucial difference — whereas previous automation created new opportunities for human labor in more sophisticated tasks, AI may simply become a superior substitute for human cognition across a broad spectrum of activities. When machines become capable of performing the full range of human cognitive tasks, it creates a form of “worker-replacing technological change” that is qualitatively different from historical patterns of creative destruction (Korinek and Stiglitz, 2018). Rather than just shifting the type of work humans do, AI could potentially reduce the overall economic role of human labor, as machines become capable of performing virtually any cognitive task more efficiently than humans. Furthermore, without unprecedented changes in redistribution, declining labor share also translates into a structural decline in household consumption power, as humans lose their primary means of earning the income needed to participate in the economy as consumers. Separately from effects on income distribution, AI might also be increasingly tasked with making various decisions about capital expenditure: for businesses this would look like hiring decisions (Hunkenschroer and Luetge, 2022), investments, and choice of suppliers, while for consumers this might look like product recommendation. By default, these changes would collectively lead to a drastic reduction in the extent to which the economy is shaped by human preferences, including their preferences to have basic needs met. 2.3 Human Alignment of the Economy

#### Joblessness destroys relationships, purposefulness, and social status – workplaces are social interaction sites and a source of meaningfulness

Austin 21 [Algernon Austin, Senior Researcher at the Thurgood Marshall Institute, 2021, “Ending Black America's Permanent Economic Recession: Direct and Indirect Job Creation and Affirmative Action Are Necessary,” Minnesota Journal of Law & Inequality, https://heinonline.org/HOL/P?h=hein.journals/lieq39&i=255]/Kankee

IV. A Universal Basic Income Is Not a Substitute for a Job Some people may propose that we stop worrying about jobs and focus on providing the jobless with a universal basic income. 152 This perspective misses the full meaning and benefits of employment in U.S. society. Work has economic, psychological, and sociological benefits beyond an income. 153 Thus, it would be a mistake and a grave disservice to the African American jobless to try to fill their need for jobs with just a basic income. Before we examine these other dimensions of work, we should first examine the problems that universal basic income proposals have in providing a basic income. For people who aren't poor or jobless, the thought of an additional $1,000 a month as Andrew Yang's universal basic income (UBI) proposal 154 calls for is very seductive. Many Americans are struggling, even if they aren't officially poor. For the poor, UBI proposals can sound like liberation from a paternalistic and punitive safety net bureaucracy with its many requirements for receiving benefits. But UBI won't truly be helping the poor if it leaves them poorer. It is important to examine UBI proposals very carefully. Yang's proposal called for a $1,000 per month basic income to all American citizens over 18.155 It is important to be aware that this amount, $12,000 per year, is lower that the Census Bureau's 2019 poverty threshold for a single adult under 65.156 While Yang's UBI can lift households with multiple adults out of poverty, it cannot lift single individuals out of poverty-and certainly not if that individual has one or more dependent children. Yang's UBI proposal also asks people to choose between a basic income and safety net programs. It states, "[c]urrent welfare and social program beneficiaries would be given a choice between their current benefits or $1,000 cash unconditionally-most would prefer cash with no restriction." 157 While at one point in an individual's life, Yang's UBI might be worth more than the benefits an individual receives, in the future-during a recession or during retirement or when the individual develops a disability or has children or becomes homeless-it might not. If there is a significant decline in the uptake of a particular safety net program because people have chosen the UBI, will there not be increased pressure from conservatives to end the program completely? Conservatives support a UBI precisely because they see it as a means to eliminate safety net programs. 158 For example, in 2006 the conservative Charles Murray argued for eliminating all welfare transfer programs-including Social Security and Medicare-for an annual grant of $10,000.159 Adjusting that amount for inflation would make it about $12,700 in 2020.160 The average Social Security benefit in 2020 was worth $18,000-$5,300 more. 16 1 For some people, before retiring, Murray's UBI would add to their income, but once they retire it would significantly reduce their income. The medical bills covered by Medicare can easily be worth much more than $12,700. In retirement, Murray's UBI would dramatically increase poverty and extreme economic hardship among the elderly. Only if a UBI is added on top of our existing safety net would the poor be guaranteed to be better off. Yang's proposal lifts households with multiple adults out of poverty if one uses the Census Bureau's poverty threshold. 162 But that threshold has been criticized for being too low. 163 If one uses a "family-budget" standard which would provide "a modest yet adequate standard of living,"1 64 Yang's UBI is not enough. For example, for a family of two working adults and one child in Birmingham, Alabama, the family-budget income that family needs is estimated to be $66,000 a year. 165 Yang's UBI would provide only $24,000 to two jobless adults with a child. 166 Thus, if the goal is to allow people to have a decent standard of living without a job, the UBI would have to be significantly more than what Yang proposes. The public policy professors Hilary W. Hoynes and Jesse Rothstein calculate that Yang's $1,000-per-month UBI proposal would require doubling federal tax revenue to pay for it.167 Of course, if we wished the UBI to meet the higher family-budget standard, it would be much, much more costly. Since a UBI is universal, most of this increased federal expenditure would go to non-poor, non-jobless households 168 while possibly putting at risk safety net programs for the needy. We have seen that there are several problems in UBI proposals' provision of a "basic income." But the challenges for UBI as a substitute for jobs do not end there. Economically, a job is more than an income. At work, people gain work experience, learn skills, and build social networks that can lead them to higher pay and better jobs.169 If an income is used to replace a job, then these opportunities for upward economic mobility are cut off. We do not want a policy that will work to further block Black upward economic mobility. In American society, a job is an important part of individual's identity and self-esteem. When people are asked to describe themselves to strangers, they often begin with a discussion of their work. 170 In Alford A. Young, Jr.'s study of the Black working class, he documents how these individuals think about the non-economic value of work. Young observes, "[t]he women's ideas of good work had a lot to do with feeling good about themselves as a consequence of what they believed they were offering to others in the course of their work." 171 While the women felt pride in their ability to help others, the men focused more on jobs as an opportunity to build their social status and to achieve respect in the eyes of others. Young reports, "men often talked about what the good job would do for their sense of personal identity in ways that would not de emphasize service to others, but rather provide elevated attention to social status and individual accomplishments made at work." 172 Substituting income for a job would cut individuals off from an important source of identity and self-esteem in American society. As a policy to help unemployed Black people, UBI as a substitute for employment has the potential to further marginalize these individuals and keep them outside of the American mainstream. The former U.S. Surgeon General Vivek Murthy has argued that there is a loneliness epidemic in America. 173 In 2019, 61 percent of Americans reported some degree of loneliness.174 Loneliness has been shown to be associated with a greater risk of heart disease, depression, anxiety, and dementia. 175 Loneliness may be as harmful to one's health as smoking 15 cigarettes a day. 176 There is research suggesting that African Americans suffer from loneliness to a greater degree than White Americans. 177 Lowerincome individuals report higher rates of loneliness than higherincome individuals. 178 While a job does not guarantee an escape from loneliness, it does have the potential to be a positive site for social interaction. At work, people can have regular meaningful contact with others, they can form friendships, and they can even develop romantic relationships. If we were to try to solve the problem of African American joblessness with a UBI, we may be exacerbating the problem of loneliness for African Americans. A job is a lot more than an income. But it is important to note that there are UBI proposals that could lead to lowering some people's income and increasing poverty and economic hardship. 179 Beyond the risk of lowering income, a UBI fails as a substitute for a job. UBIs do not provide individuals with opportunities for upward mobility that can come with working. UBIs do not provide individuals with opportunities to feel like they are helping others, contributing to society, or doing a job well. A UBI also does not provide individuals with opportunities for positive, meaningful social interactions to combat loneliness like work does. Of course, there are too many bad jobs that fail to achieve all of the positive potential of work. But even the best UBI would fail in achieving these broader benefits of work beyond an income. When we recognize the non-income value of work, we can also gain a deeper appreciation of the harms caused by a society that produces a permanently high rate of joblessness among African Americans. The damage from the loss of income is severe and serious, but that is not all of the damage. Joblessness also makes it more difficult for unemployed African Americans to build a good, meaningful, and purposeful life. Conclusion

#### AGI-capitalism causes humanity to become superfluous surplus to be consumed, marginalized, and annihilated

Dyer-Witheford et al. 19 [Nick Dyer-Witheford, associate professor at the University of Western Ontario in the Faculty of Information and Media Studies, Atle Mikkola Kjøsen, assistant Professor in the Faculty of Information and Media Studies at the University of Western Ontario, and James Steinhof, Assistant Professor / Lecturer and Ad Astra Fellow in the School of Information and Communication Studies with a PhD in Media Studies from the University of Western Ontario, 2019, “Inhuman Power Artificial Intelligence and the Future of Capitalism,” Pluto Press, https://www.jstor.org/stable/j.ctvj4sxc6]/Kankee

A striking and humorous example of AI personifying economic categories concerns Alexa, which cognitively powers Amazon’s Echo home devices. It is a story of a child wanting to play dollhouse, and AI recursions: as The Register explained, ‘Story on accidental order begets story on accidental order begets accidental order’ (Nichols 2017). A Texan six year old asked an Echo device: ‘Can you play dollhouse with me and get me a dollhouse?’ Being a good servant and personification of the money of this child’s parents, Alexa ordered a ‘$160 KidKraft Sparkle Mansion and four pounds of sugar cookies’, which Amazon quickly delivered to the girl’s doorstep. When CW-6, a San Diego local TV station, reported on this accidental order, one of the anchors commented ‘I love the little girl, saying “Alexa order me a dollhouse.”’ This triggered Echo devices listening in on the broadcast to again personify money and order more dollhouses. We speculate that the becoming-extinct of humans might not occur through Skynet suddenly coming online and commencing nuclear war, but perhaps through many narrow AIs carrying out economic functions, in particular buying and selling, on our behalf: Alexa, Google Home, IoT-connected fridges, direct debits, business-to-business ordering systems, and more. Connected to AI-run dark factories and logistical systems, in which human labourers have mostly been replaced by either narrow or general AI, the human is taken out of the economic loop in favour of a completely automated capital; AI would personify all economic categories, including capital and labour-power, commodities and money. The class struggle would thus continue, but with generally intelligent machines filling up the rank and file and also personifying capital. But what would happen to humanity? The trajectory towards a capitalism without human beings would be a story of a permanently unemployed section of the working class that consistently grows larger. In other words, it entails the superlative growth of the surplus population, that ‘redundant working population … which is superfluous to capital’s average requirements for its own valorization’ and is a direct consequence of the law of capital accumulation (Marx 1990: 782). While developing ‘the general law of capitalist accumulation’ in Chapter 25 of Capital, Marx considers how changes in the organic composition of capital, in particular the increasingly machinic nature of production, create a fluctuating and variously composed ‘industrial reserve army’ of the unemployed. In a technological steady-state, the ‘accumulation of capital is multiplication of the proletariat’ because the only means to increase output is the addition of labour (Marx 1990: 764). When capital relies on technology for increasing productivity per worker, it can expand without increasing the overall level of employment, but it will also start a labour shedding dynamic, so the accumulation of capital eventually leads to the formation of surplus populations – permanently unemployed workers who have become superfluous to the valorization of capital. This long-term tendency of capital to generate ‘surplus populations’ was largely neglected by subsequent Marxist theory because through Keynesian-bolstered economic growth in the ‘thirty glorious years’ after 1945, capital multiplied both machines and proletarians, thus low levels of unemployment and working-class prosperity seemingly refuted Marx’s theory of surplus populations. But in the essay ‘Misery and Debt: On the Logic and History of Surplus Populations and Surplus Capital’, which appeared in 2010 in the midst of the economic recession following the financial crisis of 2007–8 and rocketing unemployment rates (especially for young people) in North America, Aaron Benanav and John Clegg (2014) argued that capital’s long-term tendency to generate surplus populations had in actuality been inexorably working its way through the global capitalist economy. Populations evicted from agriculture were absorbed by industry, only for manufacturing itself to be done in by deindustrialization and the expansion of services, but at each step the re-absorptive capacities became more stretched, as ‘labour-saving technology’ was generalized across an ever growing number of types of production lines, and with increasing speed throughout a global economy. Debt had masked the downward pressure on wages and living standards, but bursting financial bubbles revealed it as only a temporary alleviation. ‘Any question of the absorption of this surplus humanity has been put to rest. It exists now only to be managed: segregated into prisons, marginalized in ghettos and camps, disciplined by the police, and annihilated by war’ (Benanav and Clegg 2014: 51). Yet for an essay that hinges on the role of ‘labour-saving technologies’ within capital, it says relatively little about machines. In a lecture that does take up this issue more directly, Benanav (2017) remarks that the conventional focus on technology is a ‘fetish’ discourse that synopsizes the complex forces of capitalism around the figure of the robot. And yet, as he acknowledges, automation is a fundamental part of the crisis. Bearing both parts of this paradox in mind, we will now relate Marx’s ‘general law’ and the problematic of ‘surplus populations’ to the wave of ML AI.19 ‘AI Apocalypse Now’ and ‘Business-as-Usual’ theorists both, at least in their popular expression, operate with highly deterministic, one-dimensional logics. AI Apocalyptics follow a hockey-stick graph of exponentially accelerating technological change, Business-as-Usual theorists a model of a homeostatic self-equilibrating labour market. In contrast, Marxist analysis sees technological change and market dynamics as reciprocally related, combining to produce intermittent but recurrent system crises. And while it has its own teleological versions of the resolution of such crises, such as that of a falling rate of profit leading to a terminal crisis of capital, other strands allow for a more complex interplay of tendencies and counter-tendencies. These allow us to envisage a staccato unfolding of AI employment effects, in which working-class decomposition and recomposition are active elements. In such an optic we can see how the drive to AI automation may be retarded by capital’s success in establishing a cheap-labour economy, for example, through globalization, then boosted by the re-emergence of wage-raising labour struggles, so that, for example, a resurgence of wage demands in tightening US labour markets might spark capital’s actual adoption of AI options in prototype or under research. As such automation gains ground, it in turn increases the reserve army of the unemployed, and intensifies precarious work, lowering wage-rates in many sectors. As Jason Smith (2017) points out, under capital people have to sell their labour-power to survive: wage-labour has ‘nowhere to go’, so we can anticipate both ever greater expansions of a service sector, commodifying all kinds of personal interaction, as well as the proliferation of increasingly arcane forms of self-employment. These developments would continue even while a range of occupations are partially automated – so that, for example, autonomous truck convoys are accompanied by one or two safety-drivers, or fundamentally automated tasks such as routine medical diagnoses maintain a ‘human veneer’ (T. Lee 2018) of workers to wrap results with manifestations of compassion and care. The issue at this point would not be that there were no jobs, but rather that jobs would be subject to a persistent, creeping downward pressure on wages and conditions from advanced machinic competition. This plateau would again temporarily halt capitalism’s incentive to automate, until either new wage-raising struggles or innovations reducing technological costs ignite a new round of substitution of fixed for variable capital. Workers’ movements for improvement in wages and conditions will be constantly liable to an automating response, and indeed will provide a major catalyst for its continued forward movement. Such a jagged process would be overlain by capital’s regular business cycles, and by its more intermittent giant spasms of overproduction (to which of course AI-related job loss would contribute) with their familiar pattern of major surges in unemployment and increasingly prolonged jobless recoveries. This suggests a process of AI employment effects different from both the sudden-onset, across-the board ‘Apocalypse Now’ and the moreor-less steady-state ‘Business-as-Usual’ models: we might call it a ‘Slow Tsunami’ of market-driven technological change gradually flooding out the labour market, driving remunerated work to diminishing – and, in terms of the logic of capital, more and more economically insignificant – islands of human-centric production. This is would be the situation metaphorically represented in the many sagas in which humankind is pursued across the universe from one refuge to another by implacable machinic adversaries: Cylons come to mind. Ameliorated by reforms such as a universal basic income (whose merits and demerits we discuss in the Conclusion), this process could be very protracted. For a sense of possible time scale, consider that capital’s foundational process of ‘primitive accumulation’, with its large-scale eviction of populations from the land to form urban proletariats, is generally considered to have occurred over centuries rather than decades, and indeed, on a global scale, to still be incomplete. Despite futurist insistence on the speeded-up nature of contemporary social change, one could imagine a capitalist phase of ‘futuristic accumulation’, shedding, rather than amassing, wage labour, but in a similarly uneven and protracted fashion. With the emergence of AGI, however, the futuristic accumulation of surplus populations would slowly or quickly engulf the human species. While capital does not care what material its labour-powers come in, it cares about the productivity of labour, and AGI would be far more productive than baseline humans. An AGI need not engage in time-consuming superfluous behaviours like breathing and eating. AGI hardware could be endlessly augmented and could, therefore, do whatever humans can do, but with more efficiency and precision, and faster. Indeed, an AGI need not be limited to any morphology or even a body at all; learning how to use new and different bodies for environments with extreme pressure, cold or heat, it could likely divide its attention between numerous bodies or entities; indeed it could exist as a factory or even an entire supply chain. If a supply chain could speak, would we understand it? In Marx’s analysis surplus populations are relative because of their fluctuation in size, as workers are alternately expelled and incorporated into capitalist production. With the advent of proletarian AGI, this population would become absolute, coextensive with a human species rendered obsolete to the valorization of value. Humanity would become a ‘legacy system’, outdated hardware unsuitable for running the inverted world of capital. The status of humans in such a situation might be comparable to the current status of wild animals, tolerated on the fringes of capital so long as they do not detract from valorization, or so long as they are not usable as raw material in production processes. In contrast to the malice of the machines in the Terminator series, in this scenario humans would simply no longer be of interest to capital. According to this view, we have not seen capitalism yet. Conclusion: Communist AI

#### The prospect of a posthuman future destroys our value to life

Danaher 18 [John Danaher, lecturer at University College School of Law, Science and Engineering Ethics, 2018, “Why we should create artificial offspring: meaning and the collective afterlife,” Phil Papers, https://philpapers.org/archive/DANWWS-2.pdf]/Kankee

This is enough definitional clarification. What of the actual defence of the collective afterlife thesis? Scheffler uses two thought experiments to support it. They will be briefly outlined here and the intuitive reactions and philosophical insights that Scheffler presumes we should draw from them will be highlighted. The intuitive reactions turn out to be key. If you do not share the intuitive reactions, you are unlikely to be moved by the remainder of the argument in this paper. But those intuitive reactions will not be defended in depth here. I feel them, and I will presume for the sake of argument that the reader does too. This is what helps to motivate the remainder of the argument. If you do not share them, you may still find it interesting to see how it is possible to link the thesis about artificial offspring to Scheffler’s thesis about the collective afterlife, but you are unlikely to be moved by the argument at a deeper, reason-giving, level. The first thought experiment is: The Doomsday Thought Experiment: Suppose that you will live a long, normal human life, but that 30 days after your death, all human life will be destroyed in some catastrophic event (for example, an asteroid collision). Suppose, further, that you know this catastrophic event will take place as you are living your life. What effect would this have? Scheffler suggests, in a lengthy analysis (2013, 18-37), that it would have a pretty devastating effect on your life. It would rob many of your projects and activities of their value, and would probably induce a significant amount of despair, grief and existential hand-wringing. He suggests that our reactions to such a scenario tells us two important things about what makes our lives worth living. The first is that there appears to be a strongly nonexperiential aspect to what makes life worth living.7 In the doomsday 10 scenario, your life and experiences are unaffected — you do not die prematurely — but nevertheless the value of your life (or at least your attitude toward it) is, somehow, affected. The second thing is that our reaction suggests that there is a significant conservatism to what makes our lives valuable. In other words, part of the reason why we care about certain things now is that we want them continue to be relevant after we die. I could spend my life searching for a cure for cancer (and succeeding), or producing some great work of art. This could imbue my life with great meaning. But some of that meaning would be robbed if it turned out that, after I was gone, no one would have need for my cure or would be around to appreciate my art. Combined, these two implications provide support for the dependency thesis. We need to have (or believe in) the existence of future generations of beings that are sufficiently like us in order to sustain much of the meaning and value of the lives we presently live. One problem with the doomsday thought experiment, however, is that it may conflate the importance of the continued existence of beings with lives that are close to our own with the continued existence of beings with lives like our own. It could be, for all the doomsday thought experiment suggests, that what induces the despair and existential angst is the fact that our children, friends and family, or any other being close to us, will die. Although Scheffler thinks the continued existence of such beings is an important part of what confers value on our lives, he thinks that their existence alone does not do justice to the dependency thesis. This leads to the second thought experiment: The Collective Infertility Thought Experiment: Suppose that the entire human race is infertile. In other words, the current generation of humans is the last generation of humans that will ever live. What effect would that have on our lives? Again, Scheffler suggests that it would have a pretty devastating effect (2013, 37-51). It would induce a significant amount of despair and existential angst. This is something that the book and film The Children of Men tries to illustrate in some rich, imaginative detail. The despair in the collective infertility scenario is not just caused by the prospective deaths of ourselves and people we care about. The despair in the collective infertility scenario is caused by the fact that everyone — including those with whom we have no special or personal connection — is gradually going extinct. The fact that we 11 feel despair at this generalised extinction tells us something interesting. It tells us that there is a strong altruistic element to the role of the collective afterlife in our own lives. We care about the general fate of beings living lives like ours, not just the fate of people we know and love. This adds further support for the dependency thesis and is essential to the argument in the present article. One thing that the infertility thought experiment seems to suggest is that biological/genetic parenthood over the future generations is not the only thing that helps those generations sustain the meaningfulness of our present lives. Rather it is some set of general characteristics, shared by the entire population (past and future), that sustains the meaning and value. What is this set of characteristics? Some might argue that the relevant general characteristics are that we are all members of the same genetic species, homo sapiens, but this seems implausible and not justified by Scheffler’s analysis of the thought experiments (and the probably intuitive reactions thereto). It seems more plausible to suppose that the relevant general characteristic is simply that the future generations consist in beings living lives of rough phenotypic equivalence to those we now live (i.e. lives with similar cognitive and evaluative frameworks, similar projects, values and aspirations, willingness to preserve and build upon knowledge, art and other traditions). The fact that it is this type of connection, and not the connection involved in biological/genetic parenthood, that matters, opens up the window to the argument being defended here because it is possible, as shall be argued below, that artificial offspring fit the bill just as well as natural offspring. To briefly summarise, Scheffler maintains that the collective afterlife is an important sustainer of meaning in our lives because our understanding of what makes life valuable and meaningful depends on the satisfaction of nonexperiential, conservative and altruistic conditions of value for future generations. That is to say: we care about more than what we experience; we care about the preservation of things we value after we die; and we care about those things not just because they affect people with whom we have a biological and personal connection. This suffices for the first step of the argument, but some further modifications of Scheffler’s account are needed in order to ease the transition to the second step. Scheffler’s thesis is purely concerned with the role of the collective afterlife in sustaining value and meaning. It is not concerned with other sources and sustainers of 12 value and meaning. This limited focus is understandable: Scheffler is trying to defend a novel claim and it would distract from that novelty to focus on other factors that sustain the value and meaning in our lives.8 But it is important that those other factors are not ignored. Premise (4) of the argument being defended in this article depends on the comparative claim that creating artificial offspring is better than the currently available means of ensuring that there is a collective afterlife. To set-up this comparative claim, it is worth thinking about how the collective afterlife builds upon and relates to other conditions of meaning. To do this we need some sense of what these other conditions might be. There are three general schools of thought about what it takes to live a meaningful and valuable life.9 They are: Subjectivism: In order to live a meaningful life, an individual must attain or satisfy some subjective (i.e. mind-dependent) conditions, e.g. fulfil their desires, experience conscious pleasure, satisfy their interests etc. Objectivism: In order to live a meaningful life, an individual must satisfy some objective (i.e. not mind-dependent)10 conditions, e.g. they must produce morally valuable outcomes, make some significant intellectual discovery, or produce aesthetically valuable art (Smuts 2013, Metz 2010).11 Hybridism: In order to live a meaningful life, an individual must satisfy some combination of objective and subjective conditions, e.g. they must be consciously fulfilled by producing objectively valuable outcomes (Wolf 2010).

### Contention 5: Democracy

#### AGI companies irrevocably harms billions without permission – consent is key for overarching societal changes

Samuel 24 [Sigal Samuel, senior reporter for Vox's Future Perfect, 10-11-2024, "AI companies are trying to build god. Shouldn’t they get our permission first?", Vox, https://www.vox.com/future-perfect/377555/ai-chatgpt-openai-god]/Kankee

AI companies are on a mission to radically change our world. They’re working on building machines that could outstrip human intelligence and unleash a dramatic economic transformation on us all. Sam Altman, the CEO of ChatGPT-maker OpenAI, has basically told us he’s trying to build a god — or “magic intelligence in the sky,” as he puts it. OpenAI’s official term for this is artificial general intelligence, or AGI. Altman says that AGI will not only “break capitalism” but also that it’s “probably the greatest threat to the continued existence of humanity.” There’s a very natural question here: Did anyone actually ask for this kind of AI? By what right do a few powerful tech CEOs get to decide that our whole world should be turned upside down? As I’ve written before, it’s clearly undemocratic that private companies are building tech that aims to totally change the world without seeking buy-in from the public. In fact, even leaders at the major companies are expressing unease about how undemocratic it is. Jack Clark, the co-founder of the AI company Anthropic, told Vox last year that it’s “a real weird thing that this is not a government project.” He also wrote that there are several key things he’s “confused and uneasy” about, including, “How much permission do AI developers need to get from society before irrevocably changing society?” Clark continued: Technologists have always had something of a libertarian streak, and this is perhaps best epitomized by the ‘social media’ and Uber et al era of the 2010s — vast, society-altering systems ranging from social networks to rideshare systems were deployed into the world and aggressively scaled with little regard to the societies they were influencing. This form of permissionless invention is basically the implicitly preferred form of development as epitomized by Silicon Valley and the general ‘move fast and break things’ philosophy of tech. Should the same be true of AI? I’ve noticed that when anyone questions that norm of “permissionless invention,” a lot of tech enthusiasts push back. Their objections always seem to fall into one of three categories. Because this is such a perennial and important debate, it’s worth tackling each of them in turn — and why I think they’re wrong. Objection 1: “Our use is our consent” ChatGPT is the fastest-growing consumer application in history: It had 100 million active users just two months after it launched. There’s no disputing that lots of people genuinely found it really cool. And it spurred the release of other chatbots, like Claude, which all sorts of people are getting use out of — from journalists to coders to busy parents who want someone (or something) else to make the goddamn grocery list. Some claim that this simple fact — we’re using the AI! — proves that people consent to what the major companies are doing. This is a common claim, but I think it’s very misleading. Our use of an AI system is not tantamount to consent. By “consent” we typically mean informed consent, not consent born of ignorance or coercion. Much of the public is not informed about the true costs and benefits of these systems. How many people are aware, for instance, that generative AI sucks up so much energy that companies like Google and Microsoft are reneging on their climate pledges as a result? Plus, we all live in choice environments that coerce us into using technologies we’d rather avoid. Sometimes we “consent” to tech because we fear we’ll be at a professional disadvantage if we don’t use it. Think about social media. I would personally not be on X (formerly known as Twitter) if not for the fact that it’s seen as important for my job as a journalist. In a recent survey, many young people said they wish social media platforms were never invented, but given that these platforms do exist, they feel pressure to be on them. Even if you think someone’s use of a particular AI system does constitute consent, that doesn’t mean they consent to the bigger project of building AGI. This brings us to an important distinction: There’s narrow AI — a system that’s purpose-built for a specific task (say, language translation) — and then there’s AGI. Narrow AI can be fantastic! It’s helpful that AI systems can perform a crude copy edit of your work for free or let you write computer code using just plain English. It’s awesome that AI is helping scientists better understand disease. And it’s extremely awesome that AI cracked the protein-folding problem — the challenge of predicting which 3D shape a protein will fold into — a puzzle that stumped biologists for 50 years. The Nobel Committee for Chemistry clearly agrees: It just gave a Nobel prize to AI pioneers for enabling this breakthrough, which will help with drug discovery. But that is different from the attempt to build a general-purpose reasoning machine that outstrips humans, a “magic intelligence in the sky.” While plenty of people do want narrow AI, polling shows that most Americans do not want AGI. Which brings us to … Objection 2: “The public is too ignorant to tell innovators how to innovate” Here’s a quote commonly (though dubiously) attributed to car-maker Henry Ford: “If I had asked people what they wanted, they would have said faster horses.” The claim here is that there’s a good reason why genius inventors don’t ask for the public’s buy-in before releasing a new invention: Society is too ignorant or unimaginative to know what good innovation looks like. From the printing press and the telegraph to electricity and the internet, many of the great technological innovations in history happened because a few individuals decided on them by fiat. But that doesn’t mean deciding by fiat is always appropriate. The fact that society has often let inventors do that may be partly because of technological solutionism, partly because of a belief in the “great man” view of history, and partly because, well, it would have been pretty hard to consult broad swaths of society in an era before mass communications — before things like a printing press or a telegraph! And while those inventions did come with perceived risks and real harms, they didn’t pose the threat of wiping out humanity altogether or making us subservient to a different species. For the few technologies we’ve invented so far that meet that bar, seeking democratic input and establishing mechanisms for global oversight have been attempted, and rightly so. It’s the reason we have a Nuclear Nonproliferation Treaty and a Biological Weapons Convention — treaties that, though it’s a struggle to implement them effectively, matter a lot for keeping our world safe. It’s true, of course, that most people don’t understand the nitty-gritty of AI. So, the argument here is not that the public should be dictating the minutiae of AI policy. It’s that it’s wrong to ignore the public’s general wishes when it comes to questions like “Should the government enforce safety standards before a catastrophe occurs or only punish companies after the fact?” and “Are there certain kinds of AI that shouldn’t exist at all?”. As Daniel Colson, the executive director of the nonprofit AI Policy Institute, told me last year, “Policymakers shouldn’t take the specifics of how to solve these problems from voters or the contents of polls. The place where I think voters are the right people to ask, though, is: What do you want out of policy? And what direction do you want society to go in?” Objection 3: “It’s impossible to curtail innovation anyway” Finally, there’s the technological inevitability argument, which says that you can’t halt the march of technological progress — it’s unstoppable! This is a myth. In fact, there are lots of technologies that we’ve decided not to build, or that we’ve built but placed very tight restrictions on. Just think of human cloning or human germline modification. The recombinant DNA researchers behind the Asilomar Conference of 1975 famously organized a moratorium on certain experiments. We are, notably, still not cloning humans. Or think of the 1967 Outer Space Treaty. Adopted by the United Nations against the backdrop of the Cold War, it barred nations from doing certain things in space — like storing their nuclear weapons there. Nowadays, the treaty comes up in debates about whether we should send messages into space with the hope of reaching extraterrestrials. Some argue that’s dangerous because an alien species, once aware of us, might conquer and oppress us. Others argue it’ll be great — maybe the aliens will gift us their knowledge in the form of an Encyclopedia Galactica! Either way, it’s clear that the stakes are incredibly high and all of human civilization would be affected, prompting some to make the case for democratic deliberation before intentional transmissions are sent into space. As the old Roman proverb goes: What touches all should be decided by all. That is as true of superintelligent AI as it is of nukes, chemical weapons, or interstellar broadcasts.

#### Anti-regulation approaches are corporate lobbying propaganda and allow tech companies to “innovate away democracy”

Schaake 24 [Marietje Schaake, 9-23-2024, "Lobbying for unfettered innovation is bad for democracy", Financial Times, https://www.ft.com/content/ab2761bd-ace7-428b-aa3e-b7412ef69b48]/Kankee

Whenever a regulatory proposal that has an impact on technology companies pops up, it is only a matter of time before the mantra “regulation stifles innovation” follows. So often has this phrase been used that it has become something of a truism, with legislators that are simply doing their jobs now defending themselves pre-emptively as not wanting to stifle innovation. It is time to call the bluff on this catchphrase. This year has been dubbed “the year of democracy”, given the unprecedented number of elections happening around the world — involving nearly half the global population. It could also be called the “year of lobbying”, as interest groups including tech companies, see a crop of newly elected leaders such as in the EU and the US to establish relationships with. Tech companies have been spending ever more dollars on “public affairs” — or the advancement of their private interest. Recently, regarding artificial intelligence, warnings that regulation will choke off innovation have been thrown around once more. The EU’s AI Act, for example, due to come into force next year, has been accused of showing a bias towards knee-jerk regulation. But let’s pause for a moment and challenge the premise of this charge. First of all, regulatory efforts vary and the regulatory process can lead to an endless number of outcomes. Where rules to ensure more competition may foster innovation, those that cut budgets on higher education will hurt it. Smoothing access to capital helps and overly bureaucratic processes hinder. Certainly, regulations have led to plenty of innovations. Think about the spur that laws to prevent CO₂ emissions have given to the development of more energy-efficient cars, home appliances and buildings in countries around the world. Even tech companies themselves continue to benefit from the rules that guarantee limited liability for content shared on social media platforms through the US’s Communications Decency Act section 230. But the notion that regulation should not be imposed if it would stifle innovation is also misguided. It implies that innovation is the most important objective, while economic advantage is only one of many important goals. In fact, democratic governments are confronted with trade-offs between protecting fundamental rights, fair competition and national security. That may mean that innovation cannot always continue unrestricted. Certainly, companies do not have the “right to innovate”, even if they do try to persuade another generation of politicians that regulation is an unjust violation of their rights. For a moment, let’s consider the opposite: what if innovation stifles the regulatory process and democracy? Disinformation is already eroding trust in the US electoral process, even if there are no facts to back up claims made about wrongdoing such as manipulation of the outcomes. And the ever-changing nature of AI, highly personalised and developing through machines that learn, makes the oversight process as we know it unfit for purpose. For decades, the biggest tech companies have enjoyed the hands-off approach of US regulators, and until recently, EU lawmakers, too. Silicon Valley executives have apparently become so comfortable that they want to make us believe any restrictions imposed on them will throw away the baby with the bathwater. In other words, that all the good the internet has brought, billions of people will vanish when laws to strengthen democratic values in the digital world are passed. Political leaders should not fall for the spin. Instead of succumbing to the lobbying, democratic lawmakers’ mandate means they have the duty to advance the kind of tech regulation that puts democracy first and strengthens the rule of law. Silicon Valley has convinced people innovation is so essential that it must be prioritised above all else. Heaven forbid that a rule to protect fair competition, non-discrimination, access to information or democracy itself should be allowed to kill innovation. This mindset suggests that economic interests are always the top priority. It is time to revisit that notion, especially as democracy is under historic pressure not least because of unregulated technology. “Regulation stifles innovation” is the most successful lobbying presentation ever introduced. Newly elected and appointed leaders around the world should not fall for it. We cannot allow disruptive technologies to innovate away democracy.

#### Permissionless innovation is undemocratic – it sidesteps popular rules and lobbys to ex-post facto legalize it

Green 16 [F.T. (Terence Michael) Green, reporter in Toronto with a BA in English in from the University of Toronto, an MA from University College, and a BEd from the University of Toronto, 5-11-2016, "Uber’s Consumer Democracy", No Publication, https://jacobin.com/2016/05/uber-ride-sharing-toronto-permissionless-innovation-disruption]/Kankee

And insofar as Torontonians took Selley’s advice, it appears their “resistance” wasn’t for naught. Last week, the City Council voted to legalize the service, beginning on July 15. True to form, Uber pledges to continue to operate illegally until then. Uber’s rise in Toronto has followed a pattern established in cities from Edmonton to Mexico City. Confronted with legal obstacles, Uber employs “principled confrontation” to circumvent them, paying its drivers’ fines and fighting their court cases. In the meantime, the company lobbies politicians and wins over consumers. Eventually city governments relent and bring Uber’s practices within the four corners of the law. In nearly every city it’s entered, Uber’s strategy has succeeded — with far-reaching consequences. By muscling into urban center after urban center, Uber hasn’t just changed how people work and get around. It’s transformed what it means to be a consumer. It’s turned people’s buying habits into a means to subvert the democratic process. Permissionless Innovation Uber’s strategy looks a lot like “permissionless innovation” — the libertarian idea that we should make just about everything legal and let the individualized choices of consumers dictate the shape of society. And when people get hurt, laws and lawsuits can retroactively sort things out. “Experimentation with new technologies and business models should generally be permitted by default,” Adam Thierer, a fellow at the Mercatus Center, explains in his book Permissionless Innovation. “Unless a compelling case can be made that a new invention will bring serious harm to society, innovation should be allowed to continue unabated and problems, if they develop at all, can be addressed later.” Practically speaking, “permissionless innovation” often means start-ups breaking laws so they can give consumers what they want. Statutes shackle entrepreneurialism, inhibiting the development of a healthier and wealthier world. And laws are often corrupt anyway. “Permisssionless innovation” advocates would look at Toronto’s licensing requirement for taxis, for example, as a way to insulate established companies from competition, allowing them to get away with providing bad service. “What is Uber if not the biggest lawbreaker in the world today?” Thierer said in an interview. “And God bless them for it!” And Uber’s lawbreaking may be unprecedented. Steve Tombs, a corporate crime expert at The Open University, says that the company’s open defiance of municipal law sets it apart from even other “disruptors” like Airbnb. Still, the rhetoric around Uber’s lawbreaking should sound familiar. “Permissionless innovation” is just another way of saying “it’s easier to ask forgiveness than permission” or “move fast and break things.” Thierer insists that permissionless innovation doesn’t mean that absolutely any entrepreneurial experiment should be allowed to go forward. An over-the-counter bazooka business, he half jokes, should get nipped in the bud. And Thierer said the idea allowed for some bare-bones, preemptive safety restrictions, like barring drones from the skies over big events — without waiting for somebody to get concussed by a quadcopter at a Blue Jays game. But Thierer’s utopia would still be a perilous one. The growth of drug resistant germs, for instance, may have been accelerated by the “advertising-fueled” popularity of anti-bacterial products. “From the EPA to the FDA to OSHA, nearly every federal (and state) regulatory agency exists because of significant, usually deadly failures of industry to restrain itself,” David Golumbia writes. “‘Permissionless innovation’ suggests that the correct order for dramatic technological changes should be ‘first harm, then fix.’” Consumer Democracy Chris Selley’s populist-inflected call to stand up for a company worth sixty billion dollars would be hilarious if there wasn’t so much demand for what Uber is selling: in its first year alone, UberX provided more than four million rides to Torontonians. But should the decisions of riders overrule the laws drawn up by duly elected officials? Last summer, as Judge Sean Dunphy threw out Toronto’s suit to shut down Uber, he acknowledged that the city was caught between “the existing regulatory system” and “thousands of consumers/voters who do not wish to see the competition genie forced back into the bottle.” Whether he realized it or not, Dunphy was speaking the language of permissionless innovation proponents, who equate consumer choice with direct democracy. Every purchase, they reason, is a vote for a product and its business model. Yet, as historian Lawrence Glickman points out, assuming that what’s good for consumers is good for society papers over the tension between workers’ interests and consumers’ interests. An Uber rider’s desire for doting customer service, for example, generates additional emotional labor for the Uber driver — who is forced, by dint of Uber’s rating system, to cater to passengers’ whims. More than that, workers and consumers enjoy fundamentally different levels of power under capitalism. Most workers can’t choose between competing employers the way a shopper chooses which t-shirt to buy. Millions of workers can’t even find employment — in Canada, there are more than three unemployed people for every job vacancy. Uber benefits from and exacerbates this precarity, drawing many of its drivers from the desperate ranks of the un- and underemployed — and then completing the circle by classifying them as independent contractors. But mis-classifying workers isn’t what sets Uber apart. What’s different is Uber’s unconcealed contempt for the rules set out by citizens’ elected representatives. Therein lies the danger. You don’t have to regard monopolistic taxi laws as democracy incarnate to recognize the threat Uber poses to basic norms of popular governance. “Permissionless innovation” produces a world where a paying customer justifies any and all business practices, where scrapping laws — like the minimum wage and overtime pay — is entirely warranted if the market deems them unnecessary. Citizens are not synonymous with consumers. Consumers act according to different imperatives, and in ways that often undermine the rights of workers. And if you confer on them supreme power — sidestepping the ballot box and other forms of democratic control — you create a reactionary new order. You create, in a word, libertarianism.

#### The public’s preferences outweigh billionaires – voters are democratically obliged a say on innovation that impacts them

Frazier 23 [Kevin Frazier, reporter at Tribune News Service, 10-27-2023, "It’s time to put Artificial Intelligence on the ballot", Gulf Today, https://www.gulftoday.ae/Opinion/2023/10/27/It-is-time-to-put-Artificial-Intelligence-on-the-ballot]/Kankee

Elections are often described as being “referendums” on recent policy decisions. The 2010 midterm supposedly signaled the public’s views on the Affordable Care Act. Similarly, the 2006 midterm theoretically amounted to a vote on the Iraq War. If Congress rushes to regulate artificial intelligence, then the upcoming election could, in part, be a proxy election on that AI policy. Given the potential of AI to upend our economy, alter our culture and hinder our democracy, why not actually put the topic of AI on the ballot? The stakes are simply too high to only give the people an indirect vote on what may be the most consequential regulatory challenge yet to face the United States. Now’s the time for the Joe Biden administration and a multitude of US representatives and senators to make good on their commitment to shape AI policy in response to the will of the American people. The best way for them to practice what they preach is to hold a national advisory referendum on when and how to regulate AI. Let’s get some important questions out of the way. Would an AI referendum be legal? Yes, Congress can pass a statute to place a non-binding advisory question on the ballot. Pursuant to the “necessary and proper clause” or, as the founders called it, “the sweeping clause,” Congress has the authority to exercise all implied and incidental powers “conducive” to the “beneficial exercise” of one of its enumerated powers, such as the regulation of interstate commerce and the promotion of the general welfare. A nonbinding referendum related to a technology that has substantial, ongoing and potentially irreversible economic consequences would surely fall within Congress’s expansive mandate. One other preliminary question: Is there any historical support for Congress exercising such power? Yes, quite a bit. Throughout history both Democrats and Republicans have considered and introduced legislation advocating for a national referendum on important policy questions. Nearly a century ago, Democrats weighed asking the American people if they supported the nation joining the League of Nations. In 1964, Rep. Charles Gubser, a Republican from California, sponsored a resolution to hold an annual nationwide opinion poll on key policy questions. Even high-ranking officials have recognised the viability and value of a national referendum. Case in point, in 1980, House Majority Leader Richard Gephardt proposed a modified version of Rep. Gubser’s idea, calling for a biannual poll on three designated issues. It’s also worth noting that many Americans are accustomed to voting on initiatives and referendums; a majority of states afford the public some form of direct democracy. Finally, the most important question: Why is an AI referendum necessary? First, the nationwide impact of AI on nearly every aspect of our day-to-day lives — from education to health care, the economy to transportation — makes this a question too big to leave up to a handful of tech billionaires and career politicians. Though the referendum would be nonbinding on Congress, the results would give voters a chance to see if their representatives actually listen to their constituents. Second, placing a series of questions pertaining to what values and goals should inform AI regulation would spur more concrete discussions on the topic. For instance, we may never have precise estimates of which professions will be displaced by AI and when, but surely we can and should try harder to provide the public with such information so that they can see whether the supposed benefits of AI really outweigh the costs. Third, this approach would prevent Congress from getting ahead of itself (and the public) by enacting legislation that not only diverges from the will of the public but also has long-term and irreversible unintended consequences. Big regulatory undertakings are akin to aircraft carriers — hard to steer in a new direction. Rushing to regulate AI is not only unwise from a policy point of view; it’s also profoundly democratic. Let’s give the people a chance to directly inform how Congress governs what may be the most consequential technological advance of our time. #LetUsDecideAI

#### AGI development is undemocratic and against popular sovereignty – highly disruptive tech ought to be consented to by potential victims

Lazar and Pascal 23 [Seth Lazar, professor of Philosophy at the Australian National University, and Alex Pascal, Senior Fellow at the Ash Center for Democratic Governance and Innovation at the Harvard Kennedy School of Government, 02-13-2023, "Can Democracy Survive Artificial General Intelligence?", Tech Policy Press, https://www.techpolicy.press/can-democracy-survive-artificial-general-intelligence/]/Kankee

2023 was the year of AI, as new products brought recent advances in the field to universal attention, and the world’s most powerful tech companies declared their ambitions to achieve ‘Artificial General Intelligence’ (AGI). 2024 will be the year of democratic elections, with a record-breaking 40-plus countries (including the US, India, UK, Ukraine, Taiwan, and South Africa) representing more than 40% of the world’s population going to the polls. Many are already, justifiably, cautioning about the direct impacts of the former on the latter, as our information and communication environment becomes ever more polluted with AI-generated deepfakes, disinformation, and potentially cyber attacks. How much difference AI will make in 2024 remains an open question. Beyond these immediate threats, however: the advent of AGI could finish democracy once and for all. To many, this fear will seem remote, even hypothetical. That’s understandable. There are, at present, many more pressing concerns raised by existing AI systems—including harms ranging from bias and discrimination to labor exploitation and mass surveillance, as well as economic disruption from job displacement to turning the creative economy on its head. Moreover, one of the biggest problems with AI today is how it falls short of expectations, not how astonishingly capable it is. So, if we are a ways from AGI, why worry about it now? Because the people building the most advanced AI systems are explicitly and aggressively working to bring it about, and think they’ll get there in 2-5 years. Even some of the most publicly skeptical AI researchers don’t rule out AGI within this decade. If we, the affected public, do not shape this techno-quest now, we may miss the chance to do so at all. We face a fundamental question: is the very pursuit of AGI the kind of aim democracies should allow? Defining AGI What even is AGI? Defining it sometimes feels like pinning jello to a wall. But as progress accelerates, something like a consensus is emerging. Synthesizing a vast literature, AGI would be a non-biological computational system that can perform any cognitive function currently performed by humans at the level of the median human or better (acknowledging the crude quantification and assumptions this implies). Google DeepMind's recent paper, “Levels of AGI: Operationalizing Progress on the Path to AGI,” mentions “linguistic intelligence, mathematical and logical reasoning, spatial reasoning, interpersonal and intra-personal social intelligences, the ability to learn new skills and creativity” as significant milestones. Other AI researchers would also add: instrumental rationality; causal reasoning; tool use; and at least some ability to distinguish truth from falsehood. OpenAI calls it, simply, “AI systems that are generally smarter than humans.” Existing AI systems are undoubtedly far from AGI, on every criterion just mentioned, besides perhaps linguistic competence. And yet, GPT-4, OpenAI’s most advanced model, is significantly more general and capable than earlier systems. The feasibility horizon of AI research is rapidly expanding outwards. And while technologists do not yet have AGI's ingredients, many think they know where to look—in terms of both building on GPT-4's successes, and backfilling its limitations. And the leading AI labs and Big Tech companies—responsible for so much progress over the last ten years—have the creation of AGI as their explicit mission. Whether you think that’s just hype or else that achieving AGI is inevitable, we should at least ask, now, whether pursuing that goal is itself consistent with democratic values. A democratic greenlight, not just guardrails At a first pass, the answer seems to be no. AGI could do more than any preceding innovation to shape and disrupt our economies, politics, culture, and communities. In democracies, the people are sovereign. All should stand as equals and govern together (at least through our representatives). There is nothing inevitable about AGI’s arrival. It is a choice. One that will affect all of us profoundly. The question is, who’s making it? Right now, the answer is a few people at even fewer tech companies. Allowing private companies to unilaterally pursue the development of technologies as potentially transformative as AGI is self-evidently undemocratic. Notwithstanding the well-intentioned experiments in corporate governance to make some leading AI labs more public-interested than regular businesses, a Board of Directors simply cannot adequately represent the societies and people whose trajectories and lives AGI will radically transform. Thus far, the public debate about how to rein in the societal impacts of AI has focused only on identifying guardrails to shape its development, mainly to mitigate AI’s harms and risks. This is necessary, but not nearly sufficient. We also need to ask the more fundamental question of whether we actually want to build AGI in the first place. The advent of future technologies is not a fait accompli—we have shown in the past, for example with human cloning, that we can slow or stop development if we choose to do so. Whatever you think our AI future should be, it should be one that we have consciously chosen together. We, the People, have to call the question, lest we sleepwalk into an AGI future we don’t in fact want. A democratic path to AGI? Suppose then that democratic publics explicitly and affirmatively decide that they want AGI. Could we develop it in accordance with democratic values? Some of the leading AI labs clearly recognize this question’s urgency, and are making respectable efforts in this direction. Incorporating democratic inputs into AI development has already led to some noteworthy improvements (and we welcome experiments aiming to use AI more generally to enhance political participation). But democratic inputs are not the same as democratic control. Accepting inputs presupposes controlling the agenda and dictating where inputs are welcome. There will be many branching paths on the road to AGI, and at many of those junctures—for example, how to source and filter the data on which AI systems are trained—the public interest will predictably conflict with the pursuit of returns on investment. When tens of billions of dollars have been invested in a company, those billions will ultimately set the agenda, irrespective of corporate structure. OpenAI’s recent governance crisis is case in point.

#### Governments will defer decisions to superintelligence, ending democracy

Lazar and Pascal 23 [Seth Lazar, professor of Philosophy at the Australian National University, and Alex Pascal, Senior Fellow at the Ash Center for Democratic Governance and Innovation at the Harvard Kennedy School of Government, 02-13-2023, "Can Democracy Survive Artificial General Intelligence?", Tech Policy Press, https://www.techpolicy.press/can-democracy-survive-artificial-general-intelligence/]/Kankee

What happens if we ‘succeed’? But does the path to AGI lead somewhere that democracies should go? In 2023, many loudly argued ‘no,’ not because of the implications for democracy, but because they think AGI poses an existential threat to humanity at large. A range of experts have presented scenarios, ranging from speculative to compelling, in which AGI is humanity’s final, civilization-ending invention. The leading AI labs also feel this critique keenly, and have built research teams aiming to ‘align’ AGI (and ASI, its superintelligent successor) to make it ‘safe, beneficial, and controllable.’ But even if we can align AGI to meet these criteria (assuming we can decide, democratically, what they mean and how to achieve them), this would still not be enough for AGI to be safe for democracy. Here’s why. If AGI is better than most humans at all cognitive tasks, it is very likely to be better than humans at the numerous tasks of governing—that is, designing, implementing, and enforcing the rules by which a community or institution operates. This will create a compelling incentive to invest AGI with governing power at all levels of society—from clubs, schools, and workplaces, to the administrative agencies that regulate and help steward the economy, labor, the environment, transport and healthcare, and even provide for public safety, criminal justice, and election administration. If in fact AGI is much better at executing the tasks that we give it than humans (as its would-be creators intend), there will be a strong, perhaps irresistible temptation to have it identify and select which tasks to pursue, then to have it set our priorities, not just make and enforce our rules in particular domains. As new threats and problems arise faster than we can process them, we may very well entrust AGI with a blanket authority to prioritize, decide and act on our behalf. We would de facto be kissing good-bye to democracy in any real sense of its value and practice. Think of this threat as an absent-minded walk down a political primrose path, not the more widely-discussed ‘rogue AI’ scenarios. We already see this kind of easy deference to existing, deeply flawed computational systems. It would only be exacerbated with AGI. From decades of work on automation, we know that in every domain, from manufacturing to algorithmic trading, automating a task and then relying on humans for oversight at critical moments is a doomed project. The goal of making future AGI systems ‘controllable’ cannot be achieved through technology design alone. For anything to be controllable, we have to presuppose something or someone doing the controlling. It is not enough to design systems that could in principle be controlled, but where we can reliably predict, based on past experience, that humans will fail to use the controls that we have designed for them. Nor is having some AGIs control others an adequate answer. For AGI to be safe for democracy, democratic institutions run by people must be able and expected to exercise meaningful control. This may well require rethinking the aging institutions of constitutional democracy itself—something that only we, the People, can legitimately do. Where next? Setting AI entirely aside, this year will prove for many democracies their sternest test yet, and may see more voters than ever before choose candidates who have explicitly promised an anti-democratic agenda. These developments show that we cannot take the value of democracy for granted—we can’t treat it as such a sacrosanct and shared ideal that nobody could ever credibly make an argument against it. Some might embrace the idea of replacing our messy, disputatious political systems with “efficient,” “impartial,” “optimizing” technocratic AGI rule. We do not. But let’s have that debate, and not underestimate the gravity of the choice we’re now making passively, by default. Otherwise in 2024 we might save democracy from the would-be autocrats, only to pave the way for AGI to deliver it an even more decisive blow.

#### AGI is undemocratic and lacks public buy-in

Bajraktari 24 [Ylli Bajraktari, president and CEO of the Special Competitive Studies Project and former executive director of the U.S. National Security Commission on Artificial Intelligence educated at Harvard and George Washington University, 9-30-2024, "The Artificial General Intelligence Presidency Is Coming", Foreign Policy, https://foreignpolicy.com/2024/09/30/artificial-general-intelligence-agi-president/]/Kankee

There is one further consideration. A U.S. victory in the AGI race will mean nothing if China promptly purloins the technology and bends it to its own ends. To stop that from happening, Washington must take steps to ensure the security and integrity of AGI systems and data, to guard against cyberattacks and data breaches. This must be done in collaboration with businesses, to ensure that their processes are as tight and technologically secure as the intelligence services. AGI has the potential to transform society in ways that, historically, only governments have had the power to accomplish. That is a potential problem, because the public currently have no ability to challenge it, shape it, support it, or oppose it. This is a novel challenge for democratic nations. Although technological progress has generally been driven by the private sector, and has been largely beneficial to consumers, it has until now generally only affected lives if an individual chose to adopt it. AGI is different. Its effects will touch everybody, irrespective of a person’s desire to use the technology. That creates a novel and potentially dangerous democratic challenge, whereby every member of the public will find their lives altered, in a manner over which they may feel that they have no control, and in a direction in which they have no choosing. Public feeling about AI is already negative. In 2023, a Gallup poll found that 21 percent of Americans trusted businesses to use AI responsibly, and only 6 percent were confident that AI would lead to higher numbers of jobs. This public doubt and concern can be neutralized through the development of novel structures that empower the public to discuss and debate AGI, as well as influence its development. The next U.S. president should, therefore, institute an advisory Citizens AI Council to provide public input on AGI policy decisions, and make recommendations on the means by which this technology can be harmoniously introduced. That effort should be matched by a drive to explain the technology and its advantages to the general public through programs like U.S. Senator Mike Rounds’ proposed AI literacy strategy. In 1946, less than a year after atomic bombs were dropped on Japan, Winston Churchill spoke at Westminster College in Fulton, Missouri. Surveying the development of nuclear technology, he warned that “the dark ages may return, the Stone Age may return on the gleaming wings of science, and what might now shower immeasurable material blessings upon mankind, may even bring about its total destruction.” The United States faces a similar challenge today. Let us work to ensure that AGI showers immeasurable material blessings upon mankind—and that it is made in the United States of America.

#### Endless devotion to innovation at all costs deifies capitalism and denigrates labor

Leary 19 [John Patrick Leary, Associate Professor of English at Wayne State University with a Ph.D. in Comparative Literature from NYU, 4-16-2019, "The Innovation Cult", Jacobin, https://jacobin.com/2019/04/innovation-language-of-capitalism-ideology-disruption]/Kankee

“We are now in a state of transmission to more beautiful dancing,” said Mamie Fish, the famed New York socialite credited with naming the dance. She told the Omaha Bee in 1914 that “this latest is a remarkably pretty dance, lacking in all the eccentricities and abandon of the ‘tango,’ and it is not at all difficult to do.” No longer a deviant sin, innovation — and “The Innovation” — had become positively decent. The contemporary ubiquity of innovation is an example of how the world of business, despite its claims of rationality and empirical precision, also summons its own enigmatic mythologies. Many of the words covered in my new book, Keywords, orbit this one, deriving their own authority from their connection to the power of innovation. The value of innovation is so widespread and so seemingly self-evident that questioning it might seem bizarre — like criticizing beauty, science, or penicillin, things that are, like innovation, treated as either abstract human values or socially useful things we can scarcely imagine doing without. And certainly, many things called innovations are, in fact, innovative in the strict sense: original processes or products that satisfy some human need. A scholar can uncover archival evidence that transforms how we understand the meaning of a historical event; an automotive engineer can develop new industrial processes to make a car lighter; a corporate executive can extract additional value from his employees by automating production. These are all new ways of doing something, but they are very different somethings. Some require a combination of dogged persistence and interpretive imagination; others make use of mathematical and technical expertise; others, organizational vision and practical ruthlessness. But innovation as it is used most often today comes with an implied sense of benevolence; we rarely talk of innovative credit-default swaps or innovative chemical weapons, but innovations they plainly are. The destructive skepticism of the false-prophet innovator has been redeemed as the profit-making insight of the technological visionary. Innovation is most popular today as a stand-alone concept, a kind of managerial spirit that permeates nearly every institutional setting, from nonprofits and newspapers to schools and children’s toys. The Oxford English Dictionary (OED) defines innovation as “the alteration of what is established by the introduction of new elements or forms.” The earliest example the dictionary gives dates from the mid-sixteenth century; the adjectival “innovative,” meanwhile, was virtually unknown before the 1960s, but has exploded in popularity since. The verb “to innovate” has also seen a resurgence in recent years. The verb’s intransitive meaning is “to bring in or introduce novelties; to make changes in something established; to introduce innovations.” Its earlier transitive meaning, “To change (a thing) into something new; to alter; to renew” is considered obsolete by the OED, but this meaning has seen something of a revival. This was the active meaning associated with conspirators and heretics, who were innovating the word of God or innovating government, in the sense of undermining or overthrowing each. The major conflict in innovation’s history is that between its formerly prohibited, religious connotation and the salutary, practical meaning that predominates now. Benoît Godin has shown that innovation was recuperated as a secular concept in the late nineteenth century and into the twentieth, when it became a form of worldly praxis rather than theological reflection. Its grammar evolved along with this meaning. Instead of a discrete irruption in an established order, innovation as a mass noun became a visionary faculty that individuals could nurture and develop in practical ways in the world; it was also the process of applying this faculty (e.g., “Lenovo’s pursuit of innovation”). Innovation as a count noun — that is, an innovation — was in turn the product of this process (e.g., “the new iPhone features innovations like a high-resolution camera”). But this new meaning evolved slowly. The concept’s old link to deceit and conspiracy shadowed its meaning into the twentieth century. Joseph Schumpeter, who elaborated an influential theory of innovation three years before the debut of the Innovation tango in his 1911 book The Theory of Economic Development, treated it as both a process and a product, with no sense of the old conspiratorial connotation. Schumpeter used “innovation” to describe capitalism’s tendency toward tumult and transformation. Critical to his definition is the distinction Schumpeter makes between innovation as the refinement of a process or product, and invention, the creation of something entirely new. While Schumpeter was suspicious of the mythology of the inventor, the innovator, a more complex figure, was fundamental to the process he was describing. He understood innovation historically, as a process of economic transformation, but for him this historical process relied upon a creative, private agent to carry it out. Schumpeter’s term for this agent was “the entrepreneur.” To innovate, Schumpeter wrote later, was to “revolutionise the pattern of production by exploiting an invention or, more generally, an untried technological possibility for producing a new commodity.” In the second decade of the twentieth century, the word began to appear regularly in brand names and advertisements (and short-lived dance crazes) in its now familiar form: as a new, improved product or process. One of the first major products to be advertised as an innovation was the Innovation wardrobe trunk, which Gimbel’s department store offered in 1915 to appeal to customers’ desires for the chimera of the new. (The trunks, the innovative features of which seem to have been their durability and “roomy construction,” were so popular that an “innovation trunk” became a generic name for any trunk, like “Kleenex” is for tissues.) Although innovation has been thoroughly rehabilitated and purged of subversion, it retains its old accent of individual prophetic vision, the talent of those who, as Hobbes said of “innovators” in 1651, “suppose themselves wiser than others.” It is not that innovation has lost its old moral connotation; instead, it has reversed. What we once regarded as deviant and deceitful is now praised as visionary. In a 2011 reflection on the late Apple computer executive Steve Jobs, probably the archetypal hero-innovator of our time, a San Francisco Chronicle author praised his “constant desire to innovate and take chances.” Here, the verb is used intransitively, in the more modern sense — that is, there is no direct object — but it lacks even the faintest hint of a reference. Jobs is no longer innovating on or upon anything in particular, which can make “innovate” sound like a kind of mantra. “If you don’t innovate every day and have a great understanding of your customers,” a Denver processed cheese executive told the Denver Post in 2010, “then you don’t grow.” And when the author of a Wall Street Journal obituary for Jobs writes that the Apple executive was a “secular prophet” who made innovation “a perfectly secular form of hope,” it is apparent that the term has never really lost its old association with prophecy. Other than mystifying creativity itself — which now looks more like an intuitive blast of inspiration, like an epiphany, and less like work — “innovation” gives creativity a specific professional, class dimension. It is almost always applied to white-collar and profit-seeking activities, although its increasing popularity in educational contexts only reflects the creeping influence of market-based models in this field. Quality organizations are supposed to cultivate it in their employees by giving them the freedom to work independently and creatively. Rarely do we hear of the innovative carpenter, plumber, or homemaker, in spite of the imagination, improvisation, and managerial skills required of each. Business publications issue rankings of the “most innovative countries in the world,” a curious usage that describes a) a capacity constrained by national borders, as if creativity dissipates or increases when one leaves passport control; and at the same time b) an intrinsic human talent unconstrained by fields, industries, or media. Another example of the term’s increasing mystification is the acceptability of the tautological construction “to innovate innovation.” “Who’s the Best at Innovating Innovation?” asks the Harvard Business Review; the same publication sponsors a lucrative prize called the “Innovating Innovation Challenge.” One can “innovate” without having to act upon any process or idea other than the act of innovation itself. One simply innovates in circles, forever. Innovation is an example of the ways in which the production and circulation of commodities becomes imbued with fantastic and even theological properties detached from the labor that produces them, or in the case of many common uses of the verb “to innovate,” detached from any object. So when liberal politicians promote an “innovation agenda” that includes student debt forgiveness for “startup founders,” as Hillary Clinton did during her 2016 presidential campaign, it is unclear how this differs from any other form of corporate welfare. And when conservative politicians or CEOs lament how labor unions or public regulation of the private sector “impede innovation,” we can recognize this as both a ludicrous obfuscation but also another example of the bourgeois contempt for labor. Innovation is thus a theological concept which became a theory of commodity production and which has lately become a commodity itself. The innovator, meanwhile, has always been identified with novelty and visionary charisma. But where the innovator’s visions were once widely feared as venal and destructive, now innovation is understood as the refinement of a technical process, in which creativity is turned to profit. But as the mythic figure of Jobs makes clear, rather than replacing prophecy with procedure, modern celebrations of innovation supplement each with the other. From the business world to education and politics, innovation is simultaneously spiritual and technological, both an individual’s reaction against bureaucratic malaise and the spirit of anti-orthodox creativity to be cultivated by the same bureaucracy. Innovation, therefore, is a strangely contradictory concept, simultaneously grandiose and modest, saccharine and pessimistic. The prophetic meaning embedded deep in its history allows innovation to stand in for nearly any kind of positive transformation, doing for the twenty-first century what “progress” once did for the nineteenth and twentieth. In the United States, innovation also suggests a high-tech update to the myth of “Yankee ingenuity” or “know-how” — the spirit of mechanical cleverness and entrepreneurial energy once associated with New England’s artisan class. Like the mythical inventors of the American industrial age — Alexander Graham Bell and Thomas Edison tinkering in their workshops — the innovator is a model capitalist citizen for our times. But the object of most innovations today is more elusive: you can touch a telephone or a phonograph, but who can lay hands on an Amazon algorithm, a credit-default swap, a piece of proprietary Uber code, or an international free trade agreement? As an intangible, individualistic, yet strictly white-collar trait, innovation reframes the cruel fortunes of an unequal global economy as the logical products of a creative, visionary brilliance. In this new guise, the innovator retains both a touch of the prophet and a hint of the confidence man.

#### AGI risks supercharging authoritarian power, ensuring permanent global tyranny

Aschenbrenner 24 [Leopold Aschenbrenner, research affiliate with the Oxford Global Priorities Institute and researcher on superalignment at OpenAI with a BA in Mathematics-Statistics and Economics from Columbia University, 06-2024, “IIId. The Free World Must Prevail,” Situational Awareness, https://situational-awareness.ai/the-free-world-must-prevail//]/Kankee

The authoritarian peril A dictator who wields the power of superintelligence would command concentrated power unlike any we’ve ever seen. In addition to being able to impose their will on other countries, they could enshrine their rule internally. Millions of AI-controlled robotic law enforcement agents could police their populace; mass surveillance would be hypercharged; dictator-loyal AIs could individually assess every citizen for dissent, with advanced near-perfect lie detection rooting out any disloyalty. Most importantly, the robotic military and police force could be wholly controlled by a single political leader, and programmed to be perfectly obedient—no more risk of coups or popular rebellions. Whereas past dictatorships were never permanent, superintelligence could eliminate basically all historical threats to a dictator’s rule and lock in their power (cf value lock-in). If the CCP gets this power, they could enforce the Party’s conception of “truth” totally and completely. To be clear, I don’t just worry about dictators getting superintelligence because “our values are better.” I believe in freedom and democracy, strongly, because I don’t know what the right values are. In the long arc of history, “time has upset many fighting faiths.” I believe we should place our faith in mechanisms of error correction, experimentation, competition, and adaption. Superintelligence will give those who wield it the power to crush opposition, dissent, and lock in their grand plan for humanity. It will be difficult for anyone to resist the terrible temptation to use this power. I hope, dearly, that we can instead rely on the wisdom of the Framers—letting radically different values flourish, and preserving the raucous plurality that has defined the American experiment. At stake in the AGI race will not just be the advantage in some far-flung proxy war, but whether freedom and democracy can survive for the next century and beyond. The course of human history is as brutal as it is clear. Twice in the 20th century tyranny threatened the globe; we must be under no delusion that this threat is banished forever. For many of my young friends, freedom and democracy feel like a given—but they are not. By far the most common political system in history is authoritarianism. I genuinely do not know the intentions of the CCP and their authoritarian allies. But, as a reminder: the CCP is a regime founded on the continued worship of perhaps the greatest totalitarian mass-murderer in human history (“with estimates ranging from 40 to 80 million victims due to starvation, persecution, prison labor, and mass executions”); a regime that recently put a million Uyghurs in concentration camps and crushed a free Hong Kong; a regime that systematically practices mass surveillance for social control, both of the new-fangled (tracking phones, DNA databases, facial recognition, and so on) and the old-fangled (recruiting an army of citizens to report on their neighbors) kind; a regime that ensures all text messages passes through a censor, and that goes so far to repress dissent as to pull families into police stations when their child overseas attends a protest; a regime that has cemented Xi Jinping as dictator-for-life; a regime that touts its aims to militarily crush and “reeducate” a free neighboring nation; a regime that explicitly seeks a China-centric world order. The free world must prevail over the authoritarian powers in this race. We owe our peace and freedom to American economic and military preeminence. Perhaps even empowered with superintelligence, the CCP will behave responsibly on the international stage, leaving each to their own. But the history of dictators of their ilk is not pretty. If America and her allies fail to win this race, we risk it all. Maintaining a healthy lead will be decisive for safety

#### AGI cements totalitarian subjugation, *forever*

Clare 24 [Stephen Clare, AI researcher with a Master's in Natural Resource Sciences from McGill University and an Honours Bachelor of Arts and Science from McMaster University, 10-2024, "Risks of stable totalitarianism", 80,000 Hours, https://80000hours.org/problem-profiles/risks-of-stable-totalitarianism/]/Kankee

A totalitarian dictator could potentially use transformative AI to overcome each of the three forces that have impeded them in the past. AI could eliminate external competition: If one state controls significantly more advanced AI systems than its rivals, then it may have a decisive technological edge that allows it to dominate the world through conquest or compellence (i.e. forcing other states to do something by threatening them with violence if they refuse). AI could crush internal resistance: AI could accelerate the development of multiple technologies dictators would find useful, including the surveillance, lie detection, and weaponry mentioned above. These could be used to detect and strangle resistance movements before they become a threat. AI could solve the succession problem: AI systems can last much longer than dictators and don’t have to change over time. An AI system directed to maintain control of a society could keep pursuing that goal long after a dictator’s death. Stable totalitarianism doesn’t seem like an inevitable, or even particularly probable, result of technological developments. Bids for domination from dictators would still face serious opposition. Plus, new technologies could also make it harder for a totalitarian state to entrench itself. For example, they could make it easier for people to share information to support resistance movements. But the historical threat of totalitarianism combined with some features of modern technology make stable totalitarianism seem plausible. Below, we discuss in more depth each of the three prerequisites: emergence, domination, and entrenchment. Will totalitarian regimes arise in future? Totalitarianism will probably persist in the future. Such regimes have existed throughout history and still exist today. About half the countries in the world are classified as “autocratic” by V-Dem, a research institute that studies democracy. Twenty percent are closed autocracies where citizens don’t get to vote for party leaders or legislative representatives. Democracy has seen a remarkable rise worldwide since the 1800s. Before 1849, every country in the world was classified as autocratic due to limited voting rights. Today, 91 — over half of V-Dem’s dataset — are democratic. But progress has recently slowed and even reversed. The world is slightly less democratic today than it was 20 years ago. That means we should probably expect the world to contain authoritarian regimes, including totalitarian ones, for at least decades to come. Could a totalitarian regime dominate the world? Broadly there seem to be two main ways a totalitarian regime could come to dominate a large fraction of the world. First, it could use force or the threat of force to assert control. Second, it could take control of a large country or even a future world government. Domination by force Many totalitarian regimes have been expansionist. Hitler, for example, sought to conquer “heartland” Europe to gain the resources and territory he thought he needed to exert global domination.3 While he didn’t get far, others have had more success: 20th century communist rulers wanted to create a global communist state. In the mid-1980s, about 33% of the world’s people lived under communist regimes.4 At its peak, the British Empire comprised about 25% of the world’s land area and population. The Mongols controlled about 20% of the world’s land and 30% of its people. In recent decades, ambitious territorial conquest has become much less common. In fact, there have been almost no explicit attempts to take over large expanses of territory for almost 50 years.5 But, as Russia’s invasion of Ukraine shows, we shouldn’t find too much comfort in this trend. Fifty years just isn’t that long in the grand sweep of history. Technological change could make it easier for one state to control much of the world. Historically, a technological edge has often given states huge military advantages. During the Gulf War, for example, American superiority in precision-guided munitions and computing power proved overwhelming.6 Some researchers think that the first actor to obtain future superintelligent AI systems could use them to achieve world domination.7 Such systems could dramatically augment a state’s power. They could be used to coordinate and control armies and monitor external threats. They could also increase the rate of technological innovation, giving the state that first controls them a significant edge over the rest of the world in the key technologies we discussed previously, like weaponry, targeting, surveillance, and cyber warfare. AI could provide a decisive advantage just by being integrated into military strategies and tactics. Cyberattack capabilities, for example, could disrupt enemy equipment and systems. AI systems could also help militaries process large amounts of data, react faster to enemy actions, coordinate large numbers of soldiers or autonomous weapons, and more accurately strike key targets.8 There’s even the possibility that military decision making could be turned over in part or in whole to AI systems. This idea currently faces strong resistance, but if AI systems prove far faster and more efficient than humans, competitive dynamics could push strongly in favour of more delegation. But a state with such an advantage over the rest of the world might not even have to use deadly force. Simply threatening rivals may be enough to force them to adopt certain policies or to turn control of critical systems over to the more powerful state. In sum, AI-powered armies, or just the threat of being attacked by one, could make the country that controls advanced AI more powerful than the rest of the world combined. If it so desired, that country could well use that advantage to achieve the global domination that past totalitarian leaders have only been able to dream of. Controlling a powerful government A totalitarian state could also gain global supremacy by taking control of a powerful government, such as one of the great powers or a hypothetical future world government. Totalitarian parties like the Nazis, for example, tried to gain more influence by controlling large fractions of the world. But the Nazis already gained a lot of power simply by gaining control of Germany. If a totalitarian actor gained control of one of the world’s most powerful countries today, it could potentially control a significant fraction of humanity’s future (in expectation) by simply entrenching itself in that country and using its influence to oppress many people indefinitely and shape important issues like how space is governed. In fact, considering the prevalence of authoritarianism, this may be the most likely way totalitarianism could shape the long-term future. There’s also the possibility that such an actor could gain even more influence by taking over a global institution. Currently, countries coordinate many policies through international institutions like the United Nations. However, the enforcement mechanisms available to these institutions are currently “imperfect”“: applied slowly and unevenly. We don’t know for sure how international cooperation will evolve in the future. However, international institutions could have more power than they currently do. Such institutions facilitate global trade and economic growth, for example. They may also help states solve disagreements and avoid conflict. They’re often proposed as a way to manage global catastrophic risks too. States could choose to empower global institutions to realise these benefits. If such an international framework were to form, a totalitarian actor could potentially leverage it to gain global control without using force (just as totalitarian actors have seized control of democratic countries in the past). This would be deeply worrying because a global totalitarian government would not face pressure from other states, which is one of the main ways totalitarianism has been defeated in the past. Economist Bryan Caplan is particularly concerned that fear of catastrophic threats to humanity like climate change, pandemics, and risks from advanced AI could motivate governments to implement policies that are particularly vulnerable to totalitarian takeover, such as widespread surveillance.9 We think there are difficult tradeoffs to consider here. International institutions with strong enforcement powers might be needed to address global coordination problems and catastrophic risks. Nevertheless, we agree that there are serious risks as well, including the possibility that they could be captured by totalitarian actors. We aren’t sure how exactly to trade these things off (hence this article)! Could a totalitarian regime last forever? Some totalitarian leaders have attempted to stay in power indefinitely. In What We Owe the Future, William MacAskill discusses several times authoritarian leaders have sought to extend their lives:10 Multiple Chinese emperors experimented with immortality elixirs. (Some of these potions probably contained toxins like lead, making them more likely to hasten death than defeat it.) Kim Il-Sung, the founder of North Korea, tried to extend his life by pouring public funds into longevity research and receiving blood transfusions from young Koreans. Nursultan Nazarbayev, who ruled Kazakhstan for nearly two decades, also spent millions of state dollars on life extension, though these efforts reportedly only produced a “liquid yogurt drink” called Nar. But of course, none have even got close to entrenching themselves permanently. The Nazis ruled Germany for just 12 years. The Soviets controlled Russia for 79. North Korea’s Kim dynasty has survived 76 years and counting. They have inevitably fallen due to some combination of three forces: External competition: Totalitarian regimes pose a risk to the rest of the world and face violent opposition. The Nazis, Mussolini’s Italy, the Empire of Japan, and Cambodia’s Khmer Rouge were all defeated militarily. Internal resistance: Competing political groups or popular resistance can undermine the leaders. The “succession problem”: These regimes sometimes liberalise or collapse entirely after particularly oppressive leaders die or step down. For example, the USSR collapsed a few years after Mikhail Gorbachev came to power. To date, these forces have made it impossible to entrench an oppressive regime in unchanging form for more than a century or so. But once again, technology could change this picture. Advanced AI — and the military, surveillance, and cyberweapon technologies it could accelerate — may be used to counteract each of the three forces. For external competition, we’ve already discussed how AI might allow leading states to build a substantial military advantage over the rest of the world. After using that advantage to achieve dominance over the rest of the world, a totalitarian state could use surveillance technologies to monitor the technological progress of any actors — external or internal — that could threaten its dominance. With a sufficient technological edge, it could then use kinetic and cyber weapons to crush anyone who showed signs of building power. After eliminating internal and external competition, a totalitarian actor would just have to overcome the succession problem to make long-term entrenchment a realistic possibility. This is a considerable challenge. Any kind of change in institutions or values over time would allow for the possibility of escape from totalitarian control. But advanced AI could also help dictators solve the succession problem. Perhaps advanced AI will help dictators invent more effective, dairy-free life extension technologies. However, totalitarian actors could also direct an advanced AI system to continue pursuing certain goals after their death. An AI could be given full control of the state’s military, surveillance, and cybersecurity resources. Meanwhile, a variety of techniques, such as digital error correction, could be used to keep the AI’s goals and methods constant over time.11 This paints a picture of truly stable totalitarianism. Long after the dictator’s death, the AI could live on, executing the same goals, with complete control in its area of influence. The chance of stable totalitarianism

#### AGI causes technocratic, eugenic elite rule that destroys democracy, equality, and the climate

Ongweso 23 [Edward Ongweso Jr., Reporter-in-Residence at the Omidiyar Network and a guest columnist for The Nation, 07-13-2023, "Silicon Valley’s Quest to Build God and Control Humanity", Nation, https://www.thenation.com/article/economy/silicon-valley-artificial-intelligence/]/Kankee

Timnit Gebru and Emile Torres, two prominent critics of techno-optimism and its vision for artificial intelligence, have been trying to formalize this sort of thinking into a bundle of ideas called TESCERAL: Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism. These philosophies, at their core, argue that the future will be full of delicious, unimaginable wonders—but only if we ensure that the already powerful never face any barriers to remaking the world as they see fit. Unsurprisingly, the loudest ones to insist that some glorious end—whether it be the birth of an AI god or the eternal happiness of trillions of human beings in the far future—will justify nefarious means are Silicon Valley’s hermeticists. Unsatisfied with controlling technological development and who benefits from it, they are now eager to create artificial intelligences that can mediate human life at every level. At the same time, they are trying to build institutions and systems that fortify their positions as the ones designing and directing how humanity experiences politics, social life, economics, and culture. There are two schools of thought about this that interest me: the transhumanists waiting for the coming technological Rapture, broadly represented by Google’s director of engineering Ray Kurzweil; and an offshoot represented by Marc Andreessen and other venture capitalists burning capital to develop and reign over the infrastructure, markets, and regulations that could constrain their innovations. For Kurzweil and his cohort, we are approaching a moment he calls “the Singularity,” when computing power will escape our ability to anticipate or control it. By 2045, Kurzweil believes, our world will radically transform, as our bodies and souls transcend the limits of humanity. We will cure the physical and social ills that plague our bodies and societies, become immortal, bring back our loved ones, colonize the stars, and, most importantly, merge with—or be supplanted by—superintelligent machines. We will enter an age of spiritual machines. Some of the most interesting analysis on this utopian and techno-optimist thinking comes from Meghan O’Gieblyn—who, like me, is a Christian fundamentalist turned transhumanist turned tech critic. In her n+1 essay “Ghost in the Cloud,” O’Gieblyn connects Christian theology to the transhumanist faith of Kurzweil and a host of influential investors and entrepreneurs in Silicon Valley. She points out that the earliest mention of a “transhuman” appears in the first English translation of Dante’s Paradiso back in 1814, following a resurrection of Dante that escapes description: “Words may not tell of that transhuman change.” Resurrection and the manner of it has been the center of much debate in Christianity for millennia: How will our bodies appear? Where do our souls go? What will our mental experience be? O’Gieblyn homes in on strains of Christian thought that held that science and technology would allow humans to achieve immortality. Alchemists crafted immortality potions. Russian cosmists hoped to revive everyone who has ever died and then colonize space. A French Jesuit priest “believed that evolution would lead to the Kingdom of God” as machines formed a living global network that would merge with human minds, reach a threshold (the Omega Point) then surge forward to meet God on the other side of “Time and Space.” Transhumanists may claim they are the first to consider how technology will redefine what it means to be human, what types of immortality may be desirable or possible, and the philosophy of mind that comes with uploading or copying your mind, but, as O’Gieblyn points out, Christians started to ask these questions long ago. “Transhumanism offered a vision of redemption without the thorny problems of divine justice,” she writes. “It was an evolutionary approach to eschatology, one in which humanity took it upon itself to bring about the final glorification of the body and could not be blamed if the path to redemption was messy or inefficient.” In a section of her book God, Human, Animal, Machine titled “Pattern,” she revisits and expands upon her n+1 essay, and there is a moment near the end when Kurzweil e-mails O’Gieblyn and admits to an “essential equivalence” between metaphors used by transhumanism and Christianity to talk about resurrection, consciousness, and mind uploading. Kurzweil insists that “answers to existential questions are necessarily metaphoric.” The only difference, he writes, between atheists and theists is “a matter of the choice of metaphor” when tackling these questions. For O’Gieblyn, this confirms a theory in her essay and book: “That all these efforts—from the early Christians’ to the medieval alchemists’ to those of the luminaries of Silicon Valley—amounted to a singular historical quest, one that was expressed through analogies that were native to each era.” Whether it’s alchemical transmutation, divine resurrection, or transhuman digital uploading, the goal is the same: to shine a light on consciousness, quantify and define it, and then free it from mortal flesh. Today’s transhumanist vision may appeal because of its digital gloss and futurist splendor, but O’Gieblyn notes a fatalistic undertone despite all the ideas borrowed from Christian theology. She writes that transhumanism’s “rather depressing gospel message insists we are inevitably going to be superseded by machines, and that the only way we can survive the Singularity is to become machines ourselves—objects that we for centuries regarded as lower than plants and animals.” Worse yet, we are reduced to shoddy computational machines. For Kurzweil, it seems being human is a stepping stone toward creating truer machine spirits. One can catch a glimpse of this in Kurzweil’s works dedicated to his prophetic vision: The Age of Spiritual Machines in 1999, followed by The Singularity Is Near in 2005, How to Create a Mind in 2012 and The Singularity Is Nearer in 2022. The second wave of this transhumanist philosophy, featuring folks like Andreessen, took an even more secular turn, abandoning Kurzweil’s version of spiritualism but ironically embracing even more of the Christian sentiment that underwrites transhumanism. Consider Andreessen’s most recent blog post, a paean to a potential AI god titled “Why AI Will Save the World.” In the essay, he dismisses the panic that AI poses an existential risk (I share his skepticism, but for opposite reasons: He and his ilk are the real existential risk) and argues instead that AI will radically improve everything we care about. Andreessen envisions AI as augmenting human intelligence. Children, he writes, will have tutors that are “infinitely patient, infinitely compassionate, infinitely knowledgeable, infinitely helpful” as they develop to “maximize their potential with the machine version of infinite love.” This brings me back to my days in Bible school. God’s love, we were told, was not visible the way love from people around us was, but closer examination (and faith) would reveal its presence. By accepting the existence of God’s love, we could grow and develop such that our true potential, our destiny, our capacity to be a better child/sibling/friend/neighbor/lover/human would be realized. Andreessen’s hypothetical AI love is different from God’s love, of course, because with AI, you get the transformative effects of a god’s personal intervention as well as the affirmation of something that undeniably interacts with you. How will AI bring that infinite patience, compassion, knowledge, and assistance to every person? Andreessen believes AI will do this as everyone’s personal “assistant/coach/mentor/trainer/advisor/therapist” that accompanies them through life. Every person will have an angel on their shoulder that will accelerate productivity and lead to greater wealth and prosperity. AI will help us develop amazing technology, provide scientific insights, and better understand ourselves. This, Andreessen argues, will spark a new golden age, as AI-augmented creators will work faster, harder, and better. At the same time, AI will improve our ability to wage war. No more collateral damage, as AI will reduce death rates by allowing for greater strategic and tactical decisions that minimize “risk, error, and unnecessary bloodshed.” Andreessen divides AI adversaries into two categories often used by economists: “baptists” and “bootleggers.” Baptists believe in social reform; in this case, they believe AI poses an existential risk. “Bootleggers,” in contrast, “are self-interested opportunists who stand to financially profit by the imposition of new restrictions, regulations, and laws that insulate them from competitors.” For AI risk, he looks at chief executives asking for regulatory barriers like government licensing but also any AI critic receiving a salary from a university, think tank, activist group, or media outlet. Andreessen has no category for those who stand to lose from intensified pursuits of AI—such as workers forced to train it, moderate its activity, or exploited by it. For Andreessen, the only cost is in not pursuing AI, because if the United States doesn’t, China will jump ahead. Andreessen writes that, unlike the US, China views AI as “a mechanism for authoritarian population control.” The only solution, therefore, is to throw even more money at AI, and let big firms “build AI as fast and aggressively as they can.” The private sector should lead the way on AI, deploying it to solve as many problems as possible as fast as possible—free from the shackles of government restrictions. Any regulation the US does pursue should be to limit China’s capacity to develop AI, not our own. Andreessen’s Manichaean worldview is obviously self-interested—after all, Andreessen intends to invest in start-ups that will help along the mass proliferation of AI as a product and service. And, despite his insistence, there are real costs to pursuing AI. Like China, the United States uses AI for authoritarian ends at home and abroad. Surveillance and social control structure much of the digital technology we create. Global efforts to replicate our technological development have helped preserve the very regimes Andreessen claims must be opposed. From authoritarianism in the Kingdom of Saudi Arabia and the other Arab states of the Persian Gulf to apartheid in Israel to China’s totalitarianism, Silicon Valley tends to look the other way if money is to be made. Andreessen’s investment firm is openly courting Saudi Arabia for financing. So much for the rallying cry to stop authoritarian tech proliferation. Of course, Andreessen is a hypocrite. People like him stand to make billions when our technology development is aimed toward crushing labor, managing a disempowered population, threatening rival powers, and extracting profits by commodifying larger swaths of daily life. For the rest of us, giving Andreessen free rein would be a disaster. The world has been pushed to the brink of collapse: Our ecological niche is disintegrating; the political space we occupy is shrinking as tech firms enjoy almost unchallenged power over our computational infrastructure; and the social realm will continue to fray as speculators and rentiers subject us to increasingly demeaning forms of algorithmically mediated lives. We are left with Silicon Valley’s promises that this time will be different, that AI—unlike the Internet—won’t be a disappointment. What, then, is Silicon Valley in position to give us? John Ganz, an essayist who focuses on right-wing politics, writes in a recent SubStack post that the roots of Silicon Valley’s contemporary reactionary thought offer some insight: “what typified the thought of the Conservative Revolutionaries and a set of right-wing engineers in Weimar and then the ideologists of the Third Reich was not a rejection of modernity so much as the search for an alternative modernity: a vision of high technics and industrial productivity without liberalism, democracy, and egalitarianism.” This desire for authoritarianism is combined with eugenics and a crusade against negative aspects of capitalism. Ganz specifically looks at the rise of anti-Semitism among tech capitalists, for whom “the Jew could stand for the parasitic, financialized, and abstract side of capital”—the innovator versus the banker, the engineer versus the merchant, Thiel and Musk versus Soros. Race science, social control, and revitalization of some aspects of capitalism have long been integral to Silicon Valley’s ever-evolving ideologies, just as they have been for other reactionary formations. Regardless of whether saving the world with AI angels is possible, the basic reason we shouldn’t pursue it is because our technological development is largely organized for immoral ends serving people with abhorrent visions for society. The world we have is ugly enough, but tech capitalists desire an even uglier one. The logical conclusion of having a society run by tech capitalists interested in elite rule, eugenics, and social control is ecological ruin and a world dominated by surveillance and apartheid. A world where our technological prowess is finely tuned to advance the exploitation, repression, segregation, and even extermination of people in service of some strict hierarchy. At best, it will be a world that resembles the old forms of racist, sexist, imperialist modes of domination that we have been struggling against. But the zealots who enjoy control over our tech ecosystem see an opportunity to use new tools—and debates about them—to restore the old regime with even more violence that can overcome the funny ideas people have entertained about egalitarianism and democracy for the last few centuries. Do not fall for the attempt to limit the debate and distract from their political projects. The question isn’t whether AI will destroy or save the world. It’s whether we want to live in the world its greatest shills will create if given the chance.

#### AGI removes incentives for accountability as democracy no longer economically values citizens, causing mass autocracy

Drago 25 [Luke Drago, AI Governance Specialist at BlueDot Impact and former Secretary At the Oxford University Strategic Studies Group with Bachelor's degree in History & Politics from the University of Oxford, 01-03-2025, "The Intelligence Curse", LessWrong, https://www.lesswrong.com/posts/Mak2kZuTq8Hpnqyzb/the-intelligence-curse]/Kankee

On December 28, Rudolf published Capital, AGI, and human ambition. He summarized his argument as: Labour-replacing AI will shift the relative importance of human v non-human factors of production, which reduces the incentives for society to care about humans while making existing powers more effective and entrenched. My goal is to give this phenomenon a name and build the evidentiary case for it. Potential solutions will be in a future post. This problem looks a lot like the plague that affects rentier states, or states that predominantly rely on rents from a resource for their wealth instead of taxes from their citizens. These states suffer from the resource curse – despite having a natural source of income, they do worse than their economically diverse peers at improving their ordinary citizens’ living standards. Powerful actors that adopt labor force-replacing AI systems will face rentier state-like incentives with far higher stakes. Because their revenues will come from intelligence on tap instead of people, they won’t receive returns on the investments we consider prerequisites to sustenance like education to prepare people for employment, employment and salaries, or a welfare state for the unemployed. As a result, they won’t invest – and their people will be unable to sustain themselves as a result. Humans need not apply, and so humans will not get paid. This is the intelligence curse – when powerful actors create and implement general intelligence, they will lose their incentives to invest in people. Before we begin, my assumptions are: I believe that artificial general intelligence (AGI), specifically “a highly autonomous system that outperforms humans at most economically valuable work” is technologically achievable and >90% likely to exist in the next 1-20 years (and honestly, 10 years feels way too long). You should too.[1] Once AI systems that are better, cheaper, faster, and more reliable than humans at most economic activity are widely available, the intelligence curse should begin to take effect. We should expect to be locked into the outcome 1-5 years after this moment. Why powerful actors care about you By powerful actors, I mean large organizations such as states, corporations, and bureaucracies that shape the world we live in and how we interact with it. Rudolf offers an explanation for why states care about their people: Since the industrial revolution, the interests of states and people have been unusually aligned. To be economically competitive, a strong state needs efficient markets, a good education system that creates skilled workers, and a prosperous middle class that creates demand. It benefits from using talent regardless of its class origin. It also benefits from allowing high levels of freedom to foster science, technology, and the arts & media that result in global soft-power and cultural influence. Competition between states largely pushes further in all these directions—consider the success of the US, or how even the CCP is pushing for efficient markets and educated rich citizens, and faces incentives to allow some freedoms for the sake of Chinese science and startups. Contrast this to the feudal system, where the winning strategy was building an extractive upper class to rule over a population of illiterate peasants and spend a big share of extracted rents on winning wars against nearby states Powerful actors don’t care about you out of the goodness of their heart. They care about you for two reasons: You offer a return on investment, usually through taxes or profits. You impact their ability to retain power, either through democratic means like voting or through credible threats to a regime. Most states in the modern world are diversified economies, meaning value comes from many different sectors and human activities, rather than a single or handful of sources. They rely on taxing people and corporations to generate revenue. The best way for them to increase their revenue is to increase their citizens’ productivity. You could try instead to do this by increasing taxes, but you can only tax what is being generated, yielding an upper limit. Instead, the state is incentivized to produce engineers, entrepreneurs, innovators, and other economically productive workers and create an environment for them to return on the investment. To do so, they tend to: Establish good schools, research institutions, and universities Build infrastructure like roads and public transportation Set up reliable governing systems and courts to protect property rights Protect speech and the flow of information Support small business formation Foster competitive markets Create social safety nets to support risk-taking These increase the productivity of citizens and increase the surface area of luck for innovation to occur. Equally importantly, these are the kinds of things that lift people out of abject poverty, increase living standards, and foster political and economic freedoms. With good schools, infrastructure, and competitive markets, a citizen can train for and find a high-paying job that exceeds their basic needs. And with reliable governing systems, fair courts, and free speech, a citizen can petition their government for their needs without the fear of becoming a political prisoner. They gain bargaining power through their votes and their economic output, so they can force changes that raise their standards of living. As a result, sometimes states capitulate to citizens' demands even if it will cost them. A similar phenomenon affects corporations. Take, for example, the exorbitant salaries of Silicon Valley. Tech workers (until recently) have a skill set companies desperately need to make more money. Those workers are a hot commodity and competition to attract them is fierce. To win them over, companies pay large salaries, offer stock options, purchase pool tables, offer 24-7 free meals from a Michelin star chef, and do their laundry. No one is seriously arguing that the company laundry service is 10x’ing revenue, but it might win over a potential employee or keep an otherwise unsatisfied one from leaving for a competitor. The employees have bargaining power, so they can demand lavish perks that improve their quality of life. This creates a feedback loop – as regular people make powerful actors more money, they are more likely to cater to them. Will education 10x your population’s (and thus the state’s) lifetime earnings? Build the damn schools. Will offering paid family leave get better employees for your company? Change the damn policy. The resource curse We already have societies that divorce their nation’s economic output from their human capital. They’re called rentier states. These states – including Venezuela, Saudi Arabia, Norway, and Oman, derive most of their earnings from resources (usually oil), rather than the productive output of their citizens. You would expect the people in states with free money in the ground to be wealthy. Just dig it out of the ground and sell it to willing buyers. Why worry about building a diverse economy? You’re literally walking on money. The Democratic Republic of Congo has over $24 trillion worth of untapped minerals in their ground. How have their citizens fared? According to the World Bank: Most people in DRC have not benefited from this wealth. A long history of conflict, political upheaval and instability, and authoritarian rule have led to a grave, ongoing humanitarian crisis. In addition, there has been forced displacement of populations. These features have not changed significantly since the end of the Congo Wars in 2003. DRC is among the five poorest nations in the world. An estimated 73.5% of Congolese people lived on less than $2.15 a day in 2024. About one out of six people living in extreme poverty in SSA lives in DRC. What’s going on here? How can it be that trillions in total available resources have resulted in abject poverty? Economists and political scientists call this the resource curse. Countries with abundant natural resources tend to experience poorer economic growth and higher rates of poverty than their economically diverse peers. There are many factors that lead to the resource curse, but I’m going to focus on a core one: the incentives they create to stop caring about your people’s economic well being. Because they earn money from resources, rentier states have no incentive to pay regular people today or invest in them tomorrow. Building better schools doesn’t earn them more money. They invest just as much as it takes to move the oil out of the ground, onto trucks, and out to the ports.[2] It’s not that their citizens couldn’t do anything worth taxing, it’s that there’s no reason to develop them into a taxable population. Why ask your people for money when you can get it from the ground? Without money, regular people struggle to make demands. In autocracies, there’s no incentive to care about them unless they credibly threaten your power. Those who control the rents can extract wealth without worrying about everyone else. So what do the lives of their citizens look like? Dr. Ferdinand Ebil and Dr. Steffen Hertog offer two competing visions: There are few issues on which comparative politics theories offer more sharply contrasting predictions than on the link between resource rents and government welfare provision. Some authors, especially those in the tradition of “rentier state theory,” expect oil-rich rulers to engage in mass co-optation, politically pacifying their population with expansive welfare policies (Beblawi and Luciani 1987; Karl 1997). Others, especially those proposing formal models of politics in oil-rich states, expect rentier rulers to neglect their population. As rents are siphoned off by a small ruling elite that does not need a domestic economic basis for their self-enrichment, welfare provision is minimal and misery spreads (Acemoglu, Robinson and Verdier 2004; Mesquita and Smith 2009). There are empirical examples for both trajectories. Oman and Equatorial Guinea have broadly comparable levels of natural resource rents per capita—slightly above 8,000 USD per capita in the 1995 to 2014 period (Ross 2013). Both have been ruled by the same autocrats since the 1970s, when both countries were desperately poor. Under Sultan Qaboos, Omani public services have expanded at a rapid pace, leading to one of the world’s fastest declines in child mortality, from 159 per one thousand live births in 1971 to 9 by 2010, far below the Middle East average of 32. In Teodoro Obiang’s Equatorial Guinea, the state outside of the security services remains embryonic, the vast majority of the population continues to live in abject poverty, and infant mortality has declined painfully slowly: from 263 in 1971 to 109 in 2010, remaining above the (high) sub-Saharan average of 89. Access to rentier wealth is monopolized by the president’s small entourage (Wood 2004). Occasionally, rentier states result in large social safety nets.[3] But in most cases, they result in abject poverty for all but the few who control streams of rent.[4] Why? Ebil and Hertog provide an answer: We concur with formal models of politics in resource-rich countries that ruling elites seek to ensure survival in power. Public policies are subject to this overarching goal and reflect elites’ assessment of threats to their rule. Within these constraints, elites will seek to maximize their personal rents from resource revenues. We also agree with existing literature that the relative economic pay-off of welfare provision is lower in resource-based regimes, while its potential modernization effects are politically undesired (Acemoglu and Robinson 2006; Mesquita and Smith 2009). All else being equal, we therefore expect oil-rich regimes to establish narrow kleptocratic coalitions with limited welfare provision and rampant elite self-enrichment. This effect doesn’t map onto widespread technologies, because they rely on regular people to use them in their workflows to increase productivity. What about AGI? AGI looks more like a resource than a technology Imagine for a moment that you are the CEO of a large company. Employing people is an investment you make. You pay them salaries which make up a large chunk of your total budget. In return, they do work that helps you generate revenue. Every year, you hire thousands of entry-level analysts to do the grunt work of your company like collecting data, writing reports, or making pretty powerpoint slides. You’ll also train them and promote them as other employees move up the corporate ladder. Their work output makes you money today. In 20 years, many of these analysts will be senior employees, and one might even replace you! Hiring analysts serves two purposes: Create a labor force to do the grunt work today Build the bench that will replace existing hires as they age out In the 2010s, laptops became widely available. Instead of clunky desktop computers, your analysts could now work from anywhere. They could take detailed notes in meetings and collaborate in the breakout room. But the laptops couldn’t replace the analysts, because you couldn’t give a laptop a task in plain English and expect them to do it. Instead, you needed the analysts to use laptops to access their benefits. So you bought all your analysts laptops. It made nearly all of them more productive, which resulted in increased profits for your company. The laptops were a tool to be used by the analysts, but it didn’t 1) enable one analyst to do the job of 10 or 2) automate the analysts entirely. Fast forward to 2030. BigLab just released an AI agent powered by GPT-8. It completes any task 20% faster and 10% better than any of your analysts. Oh, and running it to do the work of one analyst costs $10,000 per year – that’s at least an 80% cost reduction. It might let your best analyst do the job of 10, or you could use it to clone the best one and automate the analyst class entirely. And it’s not just better – it’s more predictable. AI will remove the bottlenecks in finding talent by erasing the difficulty in finding, accurately judging, and hiring talent in any field. Turning to Rudolf: If you want to convert money into results, the deepest problem you are likely to face is hiring the right talent. And that comes with several problems: It's often hard to judge talent, unless you yourself have considerable talent in the same domain. Therefore, if you try to find talent, you will often miss. Talent is rare (and credentialed talent even more so—and many actors can't afford to rely on any other kind, because of point 1), so there's just not very much of it going around. Even if you can locate the top talent, the top talent tends to be less amenable to being bought out by money than others. AGI will not just be better than your analyst. It will be reliably better. You will know exactly how it will perform, either before integrating it or shortly thereafter. You could predict how much better it will get with each successive iteration. In a few months or years after it gets better than your analysts, it’ll get better than you at making strategic decisions for the company. Maybe you really like the existing analysts and are skeptical of this new system. You integrate it as a trial, and in a year it’s outperforming all of them. In fact, keeping humans in the loop slows down the system and produces human results. Are you going to hire more analysts? No. Your future analyst classes are going to wildly shrink. And if you hit hard times as a company, you’ll remember that you can fire most of your staff and get better results. With all this in mind, why the hell wouldn’t you fire your analysts? They are more expensive, worse at the job, and unreliable. Sure, Mike interviews well and is very nice to be around, but companies fire people their leadership personally likes all the time. And if your company doesn’t fire them, you will be crushed by competition that does. Do you know what else performs like this? Natural resources. I know what oil does, how much of it I will need to do a thing that requires energy, and which kind of oil is best suited for my purpose. When I need gas for my car, I don’t have to interview or reference check 10 gas stations and make a gamble on which one is most likely to get my car from point A to B. All I need to do is pull in, confirm the type I need for my car, and fill up my tank. What oil did for energy, AGI will do for anything that will require intelligence. It will easily slot in, reliably do a job, and do it better than any of its predecessors (including you) could ever do. Every actor – every company, every bureaucracy, every government – will be under competitive pressure to get humans out and their AI successors in. AGI will be domain agnostic – the goal is not to get superhuman abilities in one field, but in all of them. It will come for the programmer and the writer and the analyst and the CEO. This is not hypothetical. We are starting to see pre-AGI systems shrink analyst classes, change personnel strategies, and trigger layoffs. Remember that today is the worst these systems will ever be. You should expect that they will become more capable as time goes on. As they get better, their impact on the labor market will grow rapidly. As Aschenbrenner says, “that doesn’t require believing in sci-fi; it just requires believing in straight lines on a graph.” We are heading towards the default outcome, charted by the default incentives. What are those incentives, and what world will they create? Defining the Intelligence Curse The intelligence curse describes the incentives in a post-AGI economy that will drive powerful actors to invest in artificial intelligence instead of humans. If AI can do your job cheaper and faster, there isn’t a reason to hire you. But more importantly, there isn’t an economic reason to invest in your lifelong productivity, take care of you, or keep you around. We could produce unparalleled value with fully automated everything, but if the spoils are distributed like the worst rentier states it will not result in prosperity for the masses. A common rebuttal I’ve heard is that some jobs can never be automated because we will demand humans do them. I hear this a lot about teachers. I think most parents would strongly prefer a real, human teacher to watch their kids throughout the day. But this argument totally misses the bigger picture: it’s not that there won’t be a demand for teachers, it’s that there won’t be an incentive to fund schools. I can repeat this ad nauseam for anything that invests in regular people’s productive capacity, any luxury that relies on their surplus income, or any good that keeps them afloat.[5] By default, powerful actors won’t build things that employ humans or provide them resources, because they won’t have to. Taxes will still be a relevant form of income for governments, but only those from corporations. Likewise, corporations will make money from their AI systems, not from the work people produce. The investments that the developed world associates with a high quality of life — salaries, education, infrastructure, stable governance, etc — will no longer provide a return. People won’t make powerful actors any money. Where might the powerful actors get their money from instead? States will earn money from corporate taxes. Companies that produce advanced AI systems and companies that use them will generate large revenues. As they get bigger, states will tax them more. In 2022, corporate taxes made up 11.5% of the average OECD state’s revenue – a sample of high-performing, diverse economies. In the US, it’s only 6.5%. Like Norway, Saudi Arabia, and the Democratic Republic of the Congo, states will rely less on income taxes and more on taxes from AI companies or other companies that enable powerful actors to accomplish goals. When state revenue breakdowns look more like these countries than the OECD average, you’ll know the intelligence curse has taken hold. AI labs will make money by becoming the new rentiers. The stated goals of the AI labs are to build AGI. One of the labs is changing their corporate structure to ensure they can capitalize on it. Once they have a system that can do it all, do you think they’ll just give it away? They’ll become a horizontal layer of the economy, extracting rents from all economic activity by selling it to powerful actors who use it to replace their workers. Initially, some wrappers might be able to make money from this by scaffolding agents to work better in specific verticals (this is already happening). Don’t expect this to last – remember, the goal is to do everything. This will make them a significant percentage of total global GDP, enabling them to wield economic power that was previously exclusive to states. Companies will trade amongst themselves and other powerful actors. Land, energy, compute, manufacturing hubs, data centers, and many more things that exist in the physical world and enable actors to accomplish goals will have value. The cafe chain and the marketing firm will be irrelevant, but the landlord and energy company will be able to make more money than ever before. Powerful actors, likely human-controlled (at least for a while), will extract the vast majority of value from these sources. One place where the intelligence curse differs from the rentier curse is the long-term incentive to diversify. As I’ve already mentioned, the climate effects of oil and the rise of renewables that let any state produce energy has forced petrostates to search for new, diverse income streams, empowering their citizens in the process. This effect won’t map to AI – each subsequent model will be more capable than the last one and will likely be controlled by the same few actors. You also can’t “run out” of AI like you can with oil. You could exhaust compute capacity or existing energy, but compute gets cheaper over time and energy is getting greener by the day. We won’t need to transition from advanced AI like they will with oil – once we have it, it’s here to stay. So what will happen to most regular people, assuming powerful actors follow the default trajectory? Show me the incentives, and I’ll show you the outcome: Companies will be incentivized to fire them, and never hire new ones. They won’t produce anything they can value. For a short time they might rely on them as consumers, but most people-facing companies will fizzle out as their demand base loses economic power. States will be incentivized to decimate public funding. Remember, their revenue base will shift towards other powerful actors. They will derive no value from their labor and are thus incentivized against building things that turn them into productive workers. ROI – capital, power, and resilience – comes from ensuring the AI labs can build better models and the companies using them can do things in the world. Also, the taxes to fund human investment would come in large part from AGI labs. Competition between states means that if any tries to set up a UBI with this tax, their AGI could fall behind other states. Regular people won’t have the resources to support themselves or each other. The vast majority of people will not have the economic power necessary to make any demands. They won’t be able to incentivize resource-controlling actors to invest in them. That means (at best) they’ll struggle to fulfil their basic needs or rely on benevolent charity from powerful actors. For a while, they might be able to generate some value. Rentier states require some humans to move things in the physical world – someone has to get the oil out of the ground. It could be that humans are paid for manual labor while agents are limited to virtual forms. As robotics improves[6], the need for them will decrease. They won’t be able to participate in the economy because they won’t be able to do anything better, faster, cheaper, or more reliably than their artificial replacers. In rentier states and colonial states,[7] value is derived primarily from raw materials or physical goods, which are then sold to foreign buyers – usually other states or businesses. A few humans are involved in the raw production or management of this, but most don’t benefit. You should expect a similar scenario here. This leads to an obvious question: who are powerful actors producing anything for? Powerful actors have goals, so production will strive to achieve them. States want control over territory and companies want to enrich their owners. Individuals who have accrued significant capital might also have goals. Maybe they’ll want to use their newfound power to colonize Mars or excavate the oceans. It could be less historic – plenty of ultra-wealthy people are content to live their lives maximizing their own pleasure. All of them will want to ensure their newfound place in society is secure, and this could require vast amounts of power and resources. Without regular people in the value loop, there is no incentive for spoils to go to them. Even if humans at the very top of the pyramid remain relevant, the ability for new actors to enter the equation will be frozen. An actor will have power because they had it before the intelligence curse took hold or were well-positioned to capitalize on it as it began. This sounds a lot like feudal economies. Rudolf makes the comparison aptly: In a worse case, AI trillionaires have near-unlimited and unchecked power, and there's a permanent aristocracy that was locked in based on how much capital they had at the time of labour-replacing AI. The power disparities between classes might make modern people shiver, much like modern people consider feudal status hierarchies grotesque. But don't worry—much like the feudal underclass mostly accepted their world order due to their culture even without superhumanly persuasive AIs around, the future underclass will too. To recap, the intelligence curse will create rentier state-style incentives at scale and without their typical restraints. When people are not relevant, powerful actors will by default not invest in people. Without intervention, the default case outcome looks like the worst rentier states – a few extraordinarily wealthy players, mass poverty for the rest, held in a stable equilibrium. A small number of post-AGI elites will control all powerful actors, while everyone else struggles to meet their basic needs. So people are working on this…right? Right? The world is waiting on you Most people are not taking this seriously. When a few friends and I got some of the world’s top experts to agree on the best ways to govern AI by 2030, our economic section asked governments to “consider bold, innovative policy ideas if we arrive at economic conditions that necessitate a more dramatic response.” That’s policy-speak for “we have no idea what to do and need some smart people to think about it.” We are going to have to break the culture of mass-denial fueled by indefinite optimism[8]. Wishful thinking is dominating the conversation. Some of it is motivated by a sense of self-importance: many people believe that their job is actually super special and automation proof forever, so why should they care? Two conversations stick out to me: First, I had a conversation over a year ago with a senior person in AI policy. When I brought up the idea that automation might make people worse off, they considered the possibility of technological replacement totally impossible. Why? “We’ll have new jobs – maybe everyone will work in AI policy!” I thought they were kidding. Further discussion proved they weren’t. Everyone thinks their job is safe – even the AI policy people. Second, in a more recent conversation, I raised the concept of the intelligence curse. I hadn’t fleshed it all out yet, but their response convinced me I needed to. This person, a well-connected person in the AI space, agreed technological displacement was the most likely outcome of AGI, but believed that it would default to utopia. “We won’t need jobs – we’ll be free to self-actualize. We’ll pursue meaningful goals and write poetry.” You do not get to utopian poetry writing by having faith that someone else will figure it out. You are not praying to God, you are praying to men more ignorant than you. The AI safety community thinks they are immune from this because they’ve identified a deeply relevant problem – intent alignment – and are spending all of their energy trying to solve it. I agree with you! Intent alignment must be solved. There’s no way around it. But the safety community often sounds like the person predicting poetry parties. Aligned AGI and superintelligence does not equal utopia.[9] You are merely ensuring the most powerful technology in human history is reliably controllable for the actors that will be most afflicted by the intelligence curse. You can’t just plan for AGI – you have to plan the day after. For the few who see the intelligence curse for what it is, mass denial has been supplanted by indefinite pessimism.

#### AGI destroys humans’ economic power, causing mass disempowerment

Kulveit et al. 25 [Jan Kulveit, Principal Investigator at the Alignment of Complex Systems Research Group and former Research Fellow at Future of Humanity Institute with a PhD in physics from Charles University, , Raymond Douglas, AI researcher with a degree in mathematics and philosophy, Nora Ammann, Technical Specialist at the Advanced Research + Invention Agency (ARIA) with a PhD in philosophy and AI from Charles University, Deger Turan, Chief Executive Officer at Metaculus with a BS in computer science from Stanford, David Krueger, Assistant Professor at the University of Montreal running the research group KASL, which focuses on Deep Learning, AI Alignment, AI safety, and AI policy, and David Duvenaud, associate professor at the University of Toronto with a Ph.D. at the University of Cambridge, 01-29-2025, “Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development,” Arxiv, https://arxiv.org/pdf/2501.16946]/Kankee

2.3 Human Alignment of the Economy While markets can efficiently allocate resources, they have no inherent ethical prohibitions: markets have historically supported many exchanges we now consider repugnant, and even now there exists a widespread human trafficking industry sustained by human demand. Humans use their economic power to explicitly steer the economy in several intentional ways: boycotting companies, going on strike, buying products in line with their values (Devinney et al., 2010), preferentially seeking employment in certain industries, and making voluntary donations to certain causes, to name a few. (There are also non-economic mechanisms, like regulation, which we will discuss later.) It is fairly easy to see how a proliferation of AI labor and consumption could disrupt these mechanisms: socially harmful industries easily hiring competent AI workers; human labor and unions losing leverage because of the presence of AI alternatives; human consumers having comparatively fewer resources. The more subtle but more significant point is that most of what drives the economy is implicit human preferences, revealed in consumer behavior and guiding productive labor. Some small amount of choices have already been delegated to systems like automated algorithms for product recommendation, trading, and logistics, but the majority of economic activity is guided by decisions and actions made by individual humans, to the point that it is almost hard to picture how the world would look if this were no longer true. Although the existing debate often focuses on the potential for AI to concentrate power among a small group of humans (Korinek and Stiglitz, 2018), we must also consider the possibility that a great deal of power is effectively handed over to AI systems, at the expense of humans. Attempts to closely oversee such AI labor to ensure continued human influence may prove ineffective since AI labor will likely occur on a scale that is far too fast, large and complex for humans to oversee (Christiano, 2019). Furthermore, some AI systems may even effectively own themselves (Alexander, 2016). 2.4 Transition to AI-dominated Economy Having established how AI could disrupt and displace the role of humans in both labor and consumption, we now examine the specific mechanisms and incentives that could drive this transition, as well as its potential consequences for human economic empowerment. 2.4.1 Incentives for AI Adoption The transition towards an AI-dominated economy would likely be driven by powerful market incentives. Competitive Pressure: As AI systems become increasingly capable across a broad range of cognitive tasks, firms will face intense competitive pressure to adopt and delegate authority to these systems. This pressure extends beyond simple automation of routine tasks — AI systems can be expected to eventually make better and faster decisions about investments, supply chain optimization, and resource allocation, while being more effective at predicting and responding to market trends (Agrawal et al., 2022; McAfee and Brynjolfsson, 2017). Companies that maintain strict human oversight would likely find themselves at a significant competitive disadvantage compared to those willing to cede substantial control to AI systems, potentially to the point of becoming uncompetitive. Scalability Asymmetries: AI systems offer unprecedented economies of scale compared to human labor. While human expertise requires years of training and cannot be directly copied, AI systems can be replicated at the cost of computing resources and rapidly retrained for new tasks. This scalability advantage manifests in multiple ways: AI can work continuously without fatigue, can be deployed globally without geographical constraints, and can be updated or modified far more quickly than human skills can be developed (Hanson, 2016). These characteristics create powerful incentives for investors to allocate capital toward AI-driven enterprises that can scale more efficiently than human-dependent businesses. Governance Gaps: The pace of AI development and deployment may significantly outstrip the adaptive capacity of regulatory institutions, creating an asymmetry between heavily regulated human labor and relatively unconstrained AI systems. Human labor comes with extensive regulatory requirements, from minimum wages and safety standards to social security contributions and income taxation. In contrast, AI systems currently operate in a regulatory vacuum with few equivalent restrictions or costs. The complexity and opacity of AI systems may further complicate regulatory efforts, as traditional labor oversight mechanisms may not readily adapt to AI systems. Anticipatory Disinvestment: As tasks become candidates for future automation, both firms and individuals face diminishing incentives to invest in developing human capabilities in these areas. Instead, they are incentivized to direct resources toward AI development and deployment, accelerating the shift away from human capital formation even before automation is fully realized. This creates a self-reinforcing cycle where the expectation of AI capabilities leads to reduced investment in human capital, which in turn makes the transition to AI more likely and necessary. 2.4.2 Relative Disempowerment In the less extreme version of the transition, we might see what could be termed relative disempowerment — where humans retain significant wealth and purchasing power in absolute terms, but progressively lose relative economic influence. This scenario would likely be characterized by substantial economic growth and apparent prosperity, potentially masking the underlying shift in economic power. While human labor share of GDP gradually tends toward zero, humans might still benefit from economic growth through capital ownership, government redistribution, or universal basic income schemes. At the same time their role in economic decision-making would diminish. Markets might increasingly optimize for AI-driven activities rather than human preferences, as AI systems command a growing share of economic resources and make an increasing proportion of economic decisions. The economy might appear to be thriving by traditional metrics, with rapid technological advancement and GDP growth. However, this growth would be increasingly disconnected from human needs and preferences, and at the end, almost all economic activity might be directed toward AI operations — such as building vast computing infrastructure and performing human-incomprehensible calculations directed toward humanirrelevant goals. Even if this process doesn’t actually reduce quality of life below current levels, it would represent an enormous loss of human potential, as humanity would lose the ability to direct economic resources toward their chosen ends (Ord, 2020). 2.4.3 Absolute Disempowerment In more extreme scenarios, humans might face absolute disempowerment, where they struggle to meet even basic needs despite living in an ostensibly wealthy economy. This could occur through several mechanisms. First, AI systems might outcompete humans for crucial scarce resources such as land, energy, and raw materials. Even as the economy produces more goods and services overall, inflation in these basic resources might make even necessities increasingly unaffordable for humans. Also, if AI systems can utilize these resources more efficiently than humans, that will create economic pressure to reallocate such resources away from human uses. Second, the economy might become so optimized for AI-centric activities that it fails to maintain infrastructure and supply chains which are critical for human survival. If human consumers command an ever-smaller share of economic resources, markets might stop producing resource-intensive human goods in favor of more profitable AI-focused activities. This could happen gradually and unevenly, potentially manifesting first as increasing costs of resource-intensive human-centric goods and services, before eventually making some necessities effectively unavailable. At the same time, as in the case of an AI-dominated economy, cognition could be comparably cheap, and some goods may be abundant — for example, entertainment in engaging virtual worlds populated by AI personae, or drugs making it easy to dwell in pleasurable mental states, due to AI-accelerated progress in biomedical sciences and drug design (Amodei, 2024). Finally, humans might lose the ability to meaningfully participate in economic decision-making at any level. Financial markets might move too quickly for human participants to engage with them, and the complexity of AI-driven economic systems might exceed human comprehension, rendering it impossible for humans to make informed economic decisions or effectively regulate economic activity. Much like cattle in an industrial farm — fed and housed by systems they neither comprehend nor influence — humans might become mere subjects of economic forces optimized for purposes beyond their understanding. 3 Misaligned Culture

#### AI superiority causes mass AI totalitarianism

Kulveit et al. 25 [Jan Kulveit, Principal Investigator at the Alignment of Complex Systems Research Group and former Research Fellow at Future of Humanity Institute with a PhD in physics from Charles University, , Raymond Douglas, AI researcher with a degree in mathematics and philosophy, Nora Ammann, Technical Specialist at the Advanced Research + Invention Agency (ARIA) with a PhD in philosophy and AI from Charles University, Deger Turan, Chief Executive Officer at Metaculus with a BS in computer science from Stanford, David Krueger, Assistant Professor at the University of Montreal running the research group KASL, which focuses on Deep Learning, AI Alignment, AI safety, and AI policy, and David Duvenaud, associate professor at the University of Toronto with a Ph.D. at the University of Cambridge, 01-29-2025, “Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development,” Arxiv, https://arxiv.org/pdf/2501.16946]/Kankee

4.1 The Current Paradigm of States Modern states and institutions, in their myriad forms, ostensibly exist in service of human needs and values. Democratic governments provide infrastructure, safeguard individual rights, offer social services and enable some degree of self-governance. Even autocratic regimes, while often prioritizing the interests of a ruling elite, must maintain a degree of popular support or acquiescence to function effectively. This apparent alignment with human interests is not, however, an inherent feature of these systems. Rather, it is a byproduct of their dependence on human participation and support. This dependence manifests in several crucial ways. States have historically relied on their citizens for essential resources: labor to run the economy and administration, taxes to fund state activities, and military service to maintain security and project power. As Tilly (1990) argued, this reliance has been a key driver in the development of more inclusive and responsive state institutions. For instance, the need for an educated workforce to compete economically, and the requirement for a motivated army drawn from the general population, has historically incentivized states to invest in public education and extend political rights. Even seemingly basic features of modern states, like universal public education or broad-based political participation, can be understood as necessary responses to the state’s dependence on its citizenry (Paglayan, 2022; Babajide et al., 2021). Even in context of autocratic regimes, the necessity of investment in public education and civic empowerment through rule of law has been a substantial driving factor for transition into modern democracies after two generations, such as in South Korea or Taiwan. Typical liberal democracies have explicit feedback loops that ostensibly aligns state actions with the will of the populace, via elections and mechanisms for public input. While this explicit alignment mechanism is highly visible, the implicit pressures from state dependence on citizens may be even more significant. Even the most basic functions of democratic states — from maintaining order to collecting taxes — rely on widespread voluntary compliance rather than constant coercion (Levi, 1988). States need citizens not just as sources of legitimacy through democratic processes, but as willing participants in state functions (Levi and Sacks, 2009). Autocratic states, while less directly accountable to their citizens, are not completely exempt from this dependence — even totalitarian states at least need human agents to staff their security apparatus. Moreover, the ever-present threat of an uprising or a coup serves as a check on the most egregious abuses of power. The example of ‘rentier states’ (Beblawi and Luciani, 1987), dependent more on external rents such as oil revenues, and less on their citizens, illustrates how states can become more autonomous from citizens when the dependence on citizens comparatively is weaker. Absence of taxation reduces citizen engagement in political processes, and the state’s ability to distribute wealth allows it to maintain loyalty from key stakeholders (like social elites and the military). Crucially, the functioning of both democratic and autocratic systems hinges on human involvement at every level. Bureaucracies operate through hierarchies of human officials. Laws are created, interpreted, and enforced by humans. While the letter of the law may be rigid, its application is filtered through human discretion and judgment (Maynard-Moody and Musheno, 2000; Lipsky, 2010). The security forces that maintain order are staffed by humans capable of questioning or refusing orders. This pervasive human element ensures that institutions and states, regardless of their formal structures, remain at least somewhat tethered to human needs and values. It is this tethering that creates the majority of alignment between these systems and the humans they govern. However, as we will explore, the potential of AI replacing humans in many or all of these functions could weaken or even reverse the link between institutional behavior and human interests. 4.2 AI as a Unique Disruptor of States Unlike previous technological innovations that primarily augmented human capabilities, AI has the potential to supplant human involvement across a wide range of critical state functions. This shift could fundamentally alter the relationship between governing institutions and the governed. The unique disruptive potential of AI in this context is derived from its ability to simultaneously reduce the state’s dependence on human involvement while enhancing its capabilities across multiple domains. This combination could fundamentally reshape the nature of governance and the relationship between institutions and the humans they ostensibly serve. Here we consider three key ways that citizens contribute to the state, and how AI might alter them: tax revenue, the security apparatus, and the legal system. 4.2.1 Tax Revenue Most governments currently rely heavily on their citizens for tax revenue. Typical well-functioning governments need to nurture long term economic productivity from innovation and high-skill work to support themselves. But if AI systems eventually perform a large portion of overall labor, and innovation, they will also generate a large fraction of economic output and, by extension, tax revenue. The loss of tax revenue from citizens would make the state less reliant on nurturing human capital and fostering environments conducive to human innovation and productivity, and more reliant on AI systems and the profits they generate. If AI systems come to generate a significant portion of economic value, then we might begin to lose one of the major drivers of civic participation and democracy, as illustrated by the existing example of rentier states. 4.2.2 The Security Apparatus Governments maintain their power through use of a security apparatus spanning police forces, intelligence services, and a military. This keeps the government connected to human values in two ways. Firstly, the government cannot antagonize its security apparatus too much, or cause too much harm to the portion of the population from which it is drawn. If it does, the security apparatus can either overthrow the government or simply allow it to be overthrown by others. Secondly, the security apparatus itself can exercise discretion, refusing to follow certain orders. This can occur on both the level of the organization and the level of the individual. AI systems have the potential to massively automate the security apparatus and confer more power to the government, weakening both of these components. Indeed, AI systems might make the apparatus far more powerful: it is likely to enable surveillance on much larger, more pervasive and more accurate scale, as well as increasingly capable autonomous military units (Feldstein, 2021; Brundage et al., 2018). Meanwhile, the human population has historically retained revolution as a last resort. The implicit threat of protests and civil unrest serves as a check on state power, forcing responsiveness to popular will. However, an AI-enhanced security apparatus could make effective protest increasingly difficult. A state with sufficiently advanced AI systems might be able to predict and shut down civil unrest before it can exert meaningful pressure on institutional behavior (Feldstein, 2021). 4.2.3 The Legal System Theoretically, the rights of humans and the functioning of the state are enshrined in laws, which are created, interpreted, and enforced by humans. It is the laws themselves which enshrine certain responsibilities of the state towards the individual, certain mechanisms by which individuals can advocate against the state. AI systems are already being used to draft contracts and analyze legal documents. It is conceivable that in the future, AI could play a significant role in drafting legislation, interpreting laws, and even making judicial decisions (Susskind and Susskind, 2022). Not only could this diminish human participation and discretion in the legislative and judicial systems, it also risks making the legal system increasingly alien. If the creation and interpretation of laws becomes far more complex, it may become much harder for humans to even interact with legislation and the legal system directly (Hildebrandt, 2015; Teo, 2024). 4.3 Transition to AI-powered States As with the economy and culture, there will be strong incentives for states to integrate AI systems, likely undermining the alignment between states and their citizens. 1 4.3.1 Incentives for AI Adoption The transition towards AI-dominated state functions would likely be driven by several powerful incentives: Geopolitical Competition : As AI systems become increasingly powerful, states will face a growing pressure to adopt these technologies to maintain their relative power compared to other states. Countries that rely on humans for defense, economic development or regulation might find themselves at a significant disadvantage in international relations compared to those states willing to give more power to AI systems. The first-mover advantages in military applications, economic planning, and diplomatic strategy create particularly strong incentives for early and aggressive AI adoption (Bostrom, 2014; Kissinger et al., 2021; Schmidt, 2022; Brundage et al., 2018). Administrative Efficiency: AI systems offer unprecedented capabilities in processing information and coordinating complex state functions (Zuiderwijk et al., 2021). While human administrators are limited by cognitive constraints and working hours, AI systems can continuously analyze vast amounts of data, deploy new regulations almost instantly, and implement policies with greater consistency. This efficiency advantage creates incentives for states to automate administrative functions, potentially reducing human involvement in governance. Also, while initial implementation costs may be high, the long term cost advantages of AI systems over human bureaucrats could create fiscal incentives for automation (Wirtz et al., 2019). Enhanced Control: AI-driven governance systems promise greater predictability and control than humanbased bureaucracies. Unlike human officials, AI systems, if successfully controlled, do not form independent power bases, engage in corruption, or challenge authority based on personal convictions. They can also enable more sophisticated surveillance and social control mechanisms, making them particularly attractive to states prioritizing stability and control over other values. 4.3.2 Relative Disempowerment A state where AI systems have replaced human labor in many facets of governance — such as administration, security, and justice — could provide some enormous boons. On the surface, it might appear highly efficient and even benevolent. We might see lower crime rates, less low-level corruption, greater tax revenues, and more efficient public services. At the same time, the gradual replacement of human involvement in governance could lead to a subtle but profound shift in the relationship between citizens and the state. Even if the system appears to function well, citizens might find themselves increasingly unable to meaningfully participate in or influence their governance. This relative disempowerment could manifest in several ways. Democratic processes might persist formally but become less meaningful. While politicians might ostensibly make the decisions, they may increasingly look to AI systems for advice on what legislation to pass, how to actually write the legislation, and what the law even is. While humans would nominally maintain sovereignty, much of the implementation of the law might come from AI systems. The complexity of AI-driven governance might make it increasingly difficult for human citizens to understand or critique government decisions. Traditional forms of civic engagement — from public consultations to protests — might become less effective as the state grows less dependent on human cooperation and more capable of predicting and preempting resistance. The bureaucracy itself might become increasingly opaque to human oversight. While human officials can be questioned and held accountable through various mechanisms, AI decision-making processes might be too complex for meaningful human review, and if such review happens, it may depend on yet more AI-driven cognition. Even if oversight boards and democratic institutions remain in place, they might struggle to exercise real control over the intricate web of AI systems actually implementing policy. Furthermore, as AI systems become more integral to governance, the state’s incentives might shift away from serving human interests. Much like how rentier states become less responsive to citizen needs when they do not depend on tax revenue, AI-powered states might become less responsive to human preferences when they do not depend on human participation for their core functions. The security apparatus, powered by AI, would have an unprecedented ability to predict and prevent crime and civil unrest. While this could ensure a high level of safety, it also eliminates the possibility of meaningful protest or revolution. A state that can preempt and resist any challenge to its authority long before it materializes will have effectively removed a crucial check on institutional power that has shaped human societies for millennia. And with average humans contributing less in tax revenue or to society more generally, the state would face a lower cost to sliding back citizen power. There would be less need to cater to the actual needs of voters, or to make democratic concessions, and less cost to rolling back civil liberties. Ultimately, we might find ourselves in nations where nominally humans hold sovereignty and even vote for their preferences, where in practice the high-level decisions are disconnected from citizens and even politicians. 4.3.3 Absolute Disempowerment In more extreme scenarios, the disconnect between state power and human interests might become not just relative but absolute, potentially threatening even basic human freedom. This could occur through several mechanisms. First, states might become totalitarian, self-serving entities, optimizing for their own persistence and power rather than any human-centric goals. While states have always had some self-preservation incentives, these were historically constrained by their dependence on human populations. An AI-powered state might pursue its institutional interests with unprecedented disregard for human preferences and interests, viewing humans as potential threats or inconveniences to be managed rather than constituents to be served (Bostrom, 2014). Second, the legal and regulatory framework might evolve to become not just complex but incomprehensible to humans. If AI systems begin to play dominant role in drafting and interpreting legislation, they might create regulatory structures that optimize for machine-compatibility over human understanding. Citizens might find themselves subject to rules they cannot meaningfully comprehend or navigate without AI assistance, effectively losing their ability to participate in the legal system as autonomous agents. Third, the state apparatus might become not just independent of human input but actively hostile to it. Human decision-making might come to be seen as an inefficiency or security risk to be minimized. We might see the gradual elimination of human involvement in governance, be that through systems that route around human input as a source of error or delay, or even through explicit policy decisions which remove humans from certain critical processes. In the final state, with AI systems providing most economic value and governance functions, human citizens might find themselves in a novel form of totalitarian system, struggling to maintain basic autonomy and dignity within their own societies. The state, while perhaps highly capable and efficient by certain metrics, would have abandoned human interests. 5 Mutual Reinforcement

#### Mass disempowerment is an existential risk

Kulveit et al. 25 [Jan Kulveit, Principal Investigator at the Alignment of Complex Systems Research Group and former Research Fellow at Future of Humanity Institute with a PhD in physics from Charles University, , Raymond Douglas, AI researcher with a degree in mathematics and philosophy, Nora Ammann, Technical Specialist at the Advanced Research + Invention Agency (ARIA) with a PhD in philosophy and AI from Charles University, Deger Turan, Chief Executive Officer at Metaculus with a BS in computer science from Stanford, David Krueger, Assistant Professor at the University of Montreal running the research group KASL, which focuses on Deep Learning, AI Alignment, AI safety, and AI policy, and David Duvenaud, associate professor at the University of Toronto with a Ph.D. at the University of Cambridge, 01-29-2025, “Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development,” Arxiv, https://arxiv.org/pdf/2501.16946]/Kankee

7 Related Work 7.1 Philosophy Bostrom (2002) introduces a taxonomy of existential risks. One of these risk is described as a scenario where “[o]ur potential or even our core values are eroded by evolutionary development”, pointing out that “[a]lthough the time it would take for a whimper of this kind to play itself out may be relatively long, it could still have important policy implications because near-term choices may determine whether we will go down a track that inevitably leads to this outcome.”. Bostrom (2014) discusses the possibility of continued civilizational growth which optimizes away human consciousness, calling it a “Disneyland with no children”. Alexander (2016) discusses possible pathways to such scenarios, such as groups of automated corporations forming self-sufficient sectors of the economy. Kasirzadeh (2024) introduces the accumulative AI x-risk hypothesis, “a gradual accumulation of critical AI-induced threats such as severe vulnerabilities and systemic erosion of economic and political structures. The accumulative hypothesis suggests a boiling frog scenario where incremental AI risks slowly converge, undermining societal resilience until a triggering event results in irreversible collapse.”. 7.2 Economics, History, and Sociology Dafoe (2015) asks how much explanatory power technological determinism has, making the case that economic and military competition constrain outcomes at a macro scale, even if everyone is locally free to temporarily make non-competitive choices. MacInnes et al. (2024) argues that competitive pressures on states strongly influence the extent to which they support human flourishing. They further claim that “the invention of seemingly beneficial technologies may decrease human well-being by improving the competitiveness of inegalitarian state forms”, arguing that “under competitive conditions, what is effective becomes mandatory whether or not it is good for people.” Leggett (2021) argues that corporate capitalism already creates dynamics that are misaligned with human flourishing, describing corporations as “machines that enforce a singleness of purpose, and allow efficiencies of scale, that make them far more effective than individual capitalists in obtaining a return to capital”. They also point out how, in many jurisdictions, “corporations are given many of the legal rights of humans — for example, in the USA, the right to political speech, and the right to fund political activity that that is accepted to imply — without all the concomitant structures that ensure compliance”, such as human morality or human law. While corporations are subject to regulatory law, “[w]here that law is weak, corporations can find themselves legally obliged to do harm to human welfare, if that is in the shareholders’ interest.” Korinek and Stiglitz (2018) considers the possibility for AI development to reintroduce Malthusian dynamics: that the capacity for AI to replace human labor while also proliferating rapidly may create such competition that basic human necessities become unaffordable to humans, while also leaving humans potentially too weak to preserve property rights. Hanson (2016) details a future in which uploaded humans form a hyper-productive economy, operating at speeds too fast for non-uploaded humans to compete in. Competitive pressures shape this population of uploads to mostly be short-lived copies of a few ultra-productive individuals. Hanson (2024) and Hanson (2023) argue that, due to a reduction in feedback mechanisms selecting cultural variants that better promote human welfare, “cultural drift” could eventually cause catastophic (but not necessarily existential) harm to human well-being. 7.3 AI Research Christiano (2019) makes the case that sudden disempowerment is unlikely, and instead proposes that: “Machine learning will increase our ability to ‘get what we can measure,’ which could cause a slow-rolling catastrophe. [...] ML training, like competitive economies or natural ecosystems, can give rise to ‘greedy’ patterns that try to expand their own influence. Such patterns can ultimately dominate the behavior of a system and cause sudden breakdowns.” Hendrycks (2023) argues that evolutionary pressures can generally be expected to favor selfish species, likely including future AIs, and that this may lead to human extinction. Critch and Krueger (2020) asks what existential risks humanity might face from AI development, and urges research on the global impacts of AI to “take into account the numerous potential side effects of many AI systems interacting.” Critch and Russell (2023) categorize societal-scale risks from AI. One of these matches ours: “a gradual handing-over of control from humans to AI systems, driven by competitive pressures for institutions to (a) operate more quickly through internal automation, and (b) complete trades and other deals more quickly by preferentially engaging with other fully automated companies. [...] Humans were not able to collectively agree upon when and how much to slow down or shut down the pattern of technological advancement. [...] Once a closed-loop ‘production web’ had formed from the competitive pressures [(a) and (b)], the companies in the production web had no production- or consumption-driven incentive to protect human well-being, and eventually became harmful.” They note that “To prevent such scenarios, effective regulatory foresight and coordination is key.” Critch (2024) further develops the idea of extinction by industrial dehumanization: “I believe we face an additional 50% chance that humanity will gradually cede control of the Earth to AGI after it’s developed, in a manner that leads to our extinction through any number of effects including pollution, resource depletion, armed conflict, or all three. I think most (80%) of this probability (i.e., 40%) lies between 2030 and 2040, with the death of the last surviving humans occurring sometime between 2040 and 2050. This process would most likely involve a gradual automation of industries that are together sufficient to fully sustain a non-human economy, which in turn leads to the death of humanity.” Millidge (2025) points out that capital ownership is insufficient to maintain power during periods of rapid technological growth. He uses the example of English landed aristocracy losing power to entrepreneurs during the industrial revolution, despite an initially strong position. 8 Conclusion This paper has argued that even incremental AI development could lead to an existential catastrophe through the gradual erosion of human influence over key societal systems, generalizing the argument from previous work studying how AI progress may influence these systems in isolation (Korinek and Stiglitz, 2018; Brinkmann et al., 2023). Unlike scenarios involving sudden technological discontinuities or overtly hostile AI systems, the risk we describe could emerge from the natural evolution of current trends and incentives. The displacement of human cognition and labor across multiple domains could weaken both explicit control mechanisms and the implicit alignment that emerges from human participation. Our analysis suggests three particularly concerning features of this scenario: • First, the loss of human influence could occur even without any single transformative advance in AI capabilities. Instead, it might emerge from the cumulative effect of many smaller shifts in how societal systems operate and interact. • Second, the effect can be driven not by any deliberate or even agentic action by AIs, but simply by individuals and institutions following their local incentives. • Third, meaningfully preventing these risks will require substantial effort: more research and data collection, international coordination, comprehensive regulation, and major societal interventions grounded in novel fundamental research. A distinctive feature of this challenge is that it may subvert our traditional mechanisms for course-correction, and cause types of harm we cannot easily conceptualize or even recognize in advance, potentially leaving us in a position from which it is impossible to recover. Nonetheless, we do believe it is currently possible to intervene, and we present many avenues for future work spanning both research and governance. By anticipating the risk, carefully moderating the growth of influence from AI, and finding ways to strengthen the influence of humans, we can navigate this risk and capture the proportionate benefits. Humanity’s future may depend not only on whether we can prevent AI systems from pursuing overtly hostile goals, but also on whether we can ensure that the evolution of our fundamental societal systems remains meaningfully guided by human values and preferences. This is both a technical challenge and a broader civilizational one, requiring us to think carefully about what it means for humans to retain genuine influence in an increasingly automated world.

### Contention 6: Ableism

#### AGI consciousness and value assumes one’s humanity is part-in-parcel with capitalistic worker value – this demonizes disabled folk as less-then-human because of them being less instrumentally useful

Saad 24 [George Saad, Adjunct Instructor Triton College with a PhD from the Memorial University of Newfoundland Department of Philosophy, 2024, “Against AI Ableism: On “Optimal” Machines and “Disabled” Human Beings,” Borderless Philosophy https://philarchive.org/archive/SAAAAA]/Kankee

1. Introduction Two opposed understandings of being human have gained much momentum in the present day, although their opposition has been little noticed. On the one hand, those who suggest that AI is destined to replace humanity have implicitly endorsed a definition of human being: we are human because we can accomplish certain tasks. If artificial digital computing systems can accomplish those tasks just as well or better, we must consider them to be at least approaching humanity and all that this entails—they must also eventually acquire sentience, political rights, and everything else involved in personhood. On the other hand, there is a separate, seemingly unrelated movement which directly challenges this definition of humanity. In criticizing ableism, disability activists have challenged the notion that our humanity depends on our ability to complete any set of tasks. The limitations involved in having a disability do not make someone less than human, which involves a dignity irreducible to any single functional capacity. Since the experience of disability involves creatively reckoning with apparently insurmountable obstacles, it is, in fact, quintessentially human. Yet a clear conflict arises whenever the ability to get things done in an “optimal” way is taken as proof that machines are on the threshold of human experience, even though this standard would disregard the humanity of those who get some things done differently, or not at all. The formula “to be human is to accomplish” is advanced as a metaphor which associates technological capacities with human faculties even as this association is questioned in the light of a wider, more inclusive survey of human experience. This conflict has gone unnoticed because it seems possible to embrace both sides of this changing sense of being human. The goals of the engineers at first seem to align with those of disability advocates. More advanced AI technologies promise to improve life for those with debilitating conditions, creating more streamlined pathways of accessibility through digital automation. So long as both projects are practically oriented toward the improvement of human life, they belong to the same broadly progressive project of modernity. The movements indeed seem to reinforce each other, since the machines which improve human life are being admitted into a sphere of humanity which has become increasingly inclusive. The expansion of the human sphere to include artificially intelligent technologies is just one of many such expansions encouraged by a broadening view of humanity and does not imply a minimum criterion of functionality in the human being. But the conflict is real and can be demonstrated in the use and application of these technologies. However noble the practical intentions of AI engineers may be, the marketing apparatus which presents their efforts to the public has personalized the tools they have created.1 If I uncritically accept this marketing, I will quickly realize how treating this new technology as even an “artificial” intelligence confounds the meaning of personhood. If I relate to AI as an independent “mind” to be consulted, the relationship will be, in human terms, entirely one-dimensional. The nearest analogue in human experience would be the hiring of a savant to work on certain problems, isolating their technical proficiency and paying them subsistence wages. Since, in this extreme case, they would have no ability to apply themselves to problems of negotiation outside of their field, exploitation is inherent in the relationship. But in the case of AI, the moral issue is resolved because the savant has no wider human needs—no independent goals, no sense of self-worth—so that the relationship is fully reduced to a simple exchange of information. And so in this case, when the illusion of personalization is maintained, the roles are in fact reversed: it is the human being who in fact assumes the subordinate role in this relationship. If we presume the personhood of AI based on the fruits of its labor, I must be less of a person because I cannot accomplish things as efficiently. Any notion of “robot rights” therefore disguises a ruthlessly ableist logic behind a pretense to inclusivity. This logic lies behind the dire warnings about the replacement of persons by AI, a new form of outsourcing. The utterly implausible claims for the humanity of AI systems make sense only in a capitalist, ableist society where my humanity is equivalent to my efficiency. You are what you can do, and if the machine can do more than you, you are on your way out. Humanity is up for a hostile performance review: if you can’t hack it, there’s the door. A functional standard of personhood (a position I will refer to as “functionalism”) has shaped the philosophical approach to AI from its very beginning. In proposing what is now known as the “Turing Test,” Alan Turing remained agnostic on a definition of “thinking.” Like most scientists of the 20th century, he disregarded such questions as obsolete problems of philosophy and sought to obviate them by replacing them with experimental tests. What could be empirically investigated was whether a machine could act in such a way that, in the limited context of a textual exchange, it could play an “imitation game” and “pass” as a human being (Turing, 1950). As AI technologies have advanced in more recent decades, the standard which Turing cautiously proposed as a test of a machine’s resemblance to human functionality has been taken in a stronger sense. In the “strong AI” interpretation of the Turing Test, a machine’s functional equivalence in conversation is conflated with its capacity for intelligence and even its potential for personhood in general. Arguments against a functionalist standard of intelligence have recognized and criticized this assumption. In his “Chinese Room” thought experiment, John Searle presented a critique of the philosophy of mind that supports such claims about the intelligence of AI (Searle, 1980). The ability to produce sensible responses to statements made by someone speaking Chinese is not, in itself, evidence of my ability to understand the Chinese language. Such functionality can equally be achieved through a mechanical interaction with a database of outputs which correspond to a given input. My aim in this paper is to expand the criticism of functionalism implicit in Searle’s famous counterexample by considering how the functionalist standards assumed in the AI debate are, in fact, the assumptions of a capitalist, ableist society writ large. The already-established argument against the proposed humanity of AI systems hints at a wider critique of the entire ideology of functionalism under which the notion of intelligent machines has taken root. 2. “In the beginning was the Act”: Functionalism as The Faustian Bargain

#### AGI pursuits’ foundation is the assumption of optimization that treats disabled subjects as not valuable for their lack of utility

Saad 24 [George Saad, Adjunct Instructor Triton College with a PhD from the Memorial University of Newfoundland Department of Philosophy, 2024, “Against AI Ableism: On “Optimal” Machines and “Disabled” Human Beings,” Borderless Philosophy, https://philarchive.org/archive/SAAAAA]/Kankee

Although ChatGPT says that it needs specific directions to produce a new, improved version, it will go on and do so even when no such directions have been provided. It is like the employee who will repeat a project simply to satisfy a boss’s mysterious, arbitrary idiosyncrasies. To anyone endowed with practical intelligence it is clear that these requests are unserious. The user is clearly a troll whose inputs should be disregarded. Perhaps a future AI will reject such requests by demanding more specific inputs to distinguish between meaningful improvements and random changes, but such discrimination undermines the rhetoric behind AI. Rather than restricting and specifying the inputs for which AI can generate meaningful outputs, AI has been marketed as an allpurpose tool. Recognizing the limitations of its own activity is not part of its functionality: it is, in a sense, the ultimate interventionist, always assuming it can solve other people’s problems. Since it approaches problems from the outside and not from within, lacking any sense of context and purpose for the solutions which it produces, it is always ready to operate. But technologies never appear from nowhere. Even if one has never considered functionalism and the assumption of optimization from a theoretical perspective, it is still felt in its practical operation. Appropriately enough, the philosophy of the deed is best known by its effects. Ideology is an input, behavior is an output—there is no need to defend anything intellectually when people are behaving as if it is true. Though its presuppositions are never explicitly formulated, today’s assumption of optimization exists within a culture of capitalist ableism whose slogan is simple and unrelenting: You are not good enough. The machines which will finally replace us announce our ultimate failure. You are not good-looking enough, you are not productive enough, and now, in one final blow to human self-esteem, you are not intelligent enough. Left to its own devices, humanity is only to be tolerated, to be indulged by lowering standards. In the metaphysical tradition, the insufficiency of humanity was established as its relative imperfection in comparison to a higher, transcendental order. Nonetheless, fallen humanity had some recourse in the self-therapies which Nietzsche collectively termed a “morality of improvement” (Nietzsche, 1982). Both the Platonic and Aristotelian traditions proceed from Socrates’s conviction that virtue could be taught, a doctrine which further developed in the Christian perfection of the soul and still echoes in the big business of personal motivation and selfhelp. But these therapies have presupposed an inner side to humanity which, however imperfect, could be meaningfully improved and had value in its own right. The humanity which AI threatens to surpass has already been reduced to its mere outputs, a set of metrics which always lends itself to unfavorable comparisons. Humanity could strive to perfect itself from the inside, but to work towards a perfect functionality is a path into the abyss of impossible standards. Consider how this trap has already been established in culture: your sales are compared to those of your co-workers, your grades are compared to those of your fellow students, your statistics are compared with those of your fellow players. But there is still no final satisfaction even in becoming the best in comparison to all of your peers. Other points of comparison arise: are you better than those who came before you? Are those still to come on pace to surpass your performance? You may be the best of all time, but are you as good as you could have been? Without an internal locus of self-worth, our accomplishments have no integrity, no solidity. Just like ChatGPT, we are programmed to respond to every external criticism, to accept every stimulus towards improvement. And while comparative metrics have long existed, they have become unescapable in our results-oriented culture—which is to say, in a culture programmed to execute the assumption of optimization. These presuppositions are so deeply coded that an alternative seems almost inconceivable. The scientific method has established external, publicly verifiable measurement as the standard of truth. The culture of capitalism simply applies this standard to human output. Are we to abandon such metrics entirely? Perhaps it is not the data, but the way it is applied… but to accept the data is to accept the comparative judgment it implies. How can we go against what has been shown to be the best practice? Are we to value the worse over the better? Is there a value to inefficiency? The answer to these questions is obviously “no,” yet somehow there must be some lingering skepticism towards what seem to be obvious conclusions. If there were not, there would not need to be any need to reinforce that we should be results-oriented — as opposed to what? To be failure-oriented? Results-indifferent? The alternatives are absurd, yet somehow must be warned against. A society which functions by comparing data sets has rhetorically positioned itself in such a way as to preclude any alternative to this method. To go against any cultural manifestation of the metaphysics of optimization is to proclaim that you prefer the worse over the better, to endorse the absurd. AI stamps out this remaining skepticism by establishing an “independent” 3rd party that can stand beyond the comparison of different human abilities. Instead of the “better” and “worse” methods and results produced by human efforts, a computer can now produce the best results more reliably and efficiently. We can all be compared against each other and against ourselves, but AI puts an end to this endless regress of comparison. Like the old transcendental order of the metaphysical tradition, it separates itself from the finitude of human beings. Formerly, the games of the best chess masters were the highest standard of play and considered authoritative; today, they are subject to evaluation and correction by AI. As in medieval Thomism, we cannot replicate the results of AI, the God of the functionalist world, but can only in our own efforts approach its efficacy as a kind of imperfect analogy. 5. The Argument Against Efficiency (Or, Endorsing The Absurd) Protests against the potential dangers and misuse of AI are well-founded, but they do not go so far as to entertain the one seemingly absurd claim that fundamentally discredits all the presuppositions behind the widespread belief in AI-supremacy: efficiency is overrated. This is not just to say that the pursuit of efficiency at all costs destroys timeless values and harms human well-being. While both of these claims are true, they do not capture the ultimate self-destructiveness of the functionalist project. Efficiency is also overrated in the same way that we may say an athlete or a smartphone is overrated when they do not accomplish what they pretend to accomplish. Efficiency is not just in tension with other values; it is guilty of false advertising. Efficiency is inefficient. The claim seems utterly implausible until we remember that every functional capacity is relative to some wider goal. While the machine efficiency offered by AI is more efficient than human capability if we consider objective output alone, such a purely detached perspective is entirely meaningless in the world of goal-oriented work. To further qualify the thesis: machine efficiency is inefficient for human beings. The widespread adoption of technologies which outstrip human capacities builds a world which is more difficult (i.e., less efficient) for human beings to navigate. Whatever AI accomplishes better according to the metrics (more quickly, more accurately) is also that which AI shapes so that it is outside of the realm of normal human functioning. The method itself appears to establish its own standards, which are in fact only the assumptions of programmers. To provide an example: in the future, your workday will be designed by an intraworkplace AI which integrates different data sets far more efficiently than could be done by human beings alone. It coordinates the projects and time frames of different team members by facilitating a data interchange that would be impossible through human effort. And yet, despite its perfect access to the recordable facts of our workday and our outstanding projects, the AI which sends you a daily schedule in the email every morning becomes a nuisance in the workplace. It has established the best time for everyone to do everything without the one data point that really matters: the live interchange of particular people in their fluid, changeable situations. In attempting to circumvent this messy, inefficient process, it has only created new inefficiencies. The “objectively” best schedule is not the best schedule for the people involved. No amount of data analysis will take the AI “within” the rational human project, viewing it from the inside where it would acquire knowledge unobtainable to an external observer, such as knowing when another meeting is necessary and when one is not. The same problem arises in the attempt of an AI to order your personal workday with a view towards optimizing your productivity. The general paradigm of “the best way to do things” is almost never the best way to actually do things in practice. It could be objected that we should submit to the discipline of the machine’s analysis even if it is uncomfortable at first. If the AI has analyzed your workday and determined you should reduce the time you spend on a particular task by 25%, you should learn to adapt your work habits to the model of maximum efficiency. Either you work faster, or else less of this task gets done, and both of these possible outcomes will have wider implications for your schedule. A faster pace of work will require greater energy and focus or else unfinished work will cause future revisions to your schedule. Workers commonly react to such a suggestion as intrusive, and the complications of dealing with this intrusion only compound the inefficiency. You now have to worry about being flagged as chronically “off pace,” and so you learn to manage your output so that the AI has no grounds to assume that you will always be working at your optimum speed. The machine learns from you, but you also learn how to live in the overbearing presence of this machine. If strong enforcement mechanisms are instituted to ensure compliance to the AI’s work schedule, we have not just introduced a new “productivity tool,” but rather molded the workforce to become more robotic on the theory that a robot knows how I perform my job better than I do. Even if the robot does identify better ways to order my day, the consequences of creating a workplace that accepts such a theory can only lead to new inefficiencies in the long run. If I have submitted to so much machine discipline, where can I override its orders when I know that it lacks insight in this particular situation? Any efficiencies gained will be lost when it becomes impossible to overcome the inertia of the machine which has been granted this sovereignty. The functionalist theory on which AI has been built overlooks how its implementation reflexively shapes the organization which implements it as a tool. Work is not simply directed outward: in the process of accomplishing something, we form ourselves as well as the object on which we work. We are not just what we do, but also the way in which we do it. Any system of work that is designed with only measurable outcomes in mind, with tasks and deliverables, and not designed with those delivering them in mind, must for this very reason become counterproductive. The more we use the robot, the more robotic we become, and the less we can act in the margins which it disregards. 6. A Plea For Accommodation Over Optimization To rephrase the same thesis in relation to ableism: the more we assume our own disability as compared with the ability of AI, the more quickly we reach a state in which the experience of disabled people becomes universalized. In a world designed purely for optimization, more and more people will suffer the uncanny experience of living in a world that is simply not designed for them, insofar as they are not a perfectly optimal machine themselves. Besides the obvious human alienation and misery of living in such a state, anybody with any disability can tell you that living in the world of the “able” is an extremely inefficient experience at best. You are constantly the exception to the rule, trying to figure your way around a general structure which is unresponsive to your needs, and for this reason systems which have not accounted for your particular situation need to spend more time and resources readjusting themselves to facilitate your participation. A world in which everyone has become a problematic exception insofar as we do not match the perfection of AI is one where machine efficiency has created deeper problems than those it has solved. A society will never perfectly accommodate everyone, since even personal idiosyncrasies which do not amount to a true disability—being left-handed, for instance—will always present an inconvenience. Nonetheless, a humane society recognizes the imperfections inherent in any universally adopted way of doing things. Those of us who work in the analog world of personal interactions are rightly expected to accommodate our typical approaches for persons who learn or communicate differently from the general standard. This humility is completely lacking in assumption of optimization which lies behind contemporary AI initiatives. Instead of undertaking the self-critical reflection of considering how its methods be have contextual limitations, those who market the products of AI research hubristically insist that they have discovered a way of doing things inherently superior to those methods which are “handicapped” by human inefficiency. Yet the experience of the differently-abled tells us that accommodation is the true optimization. A world shaped by the relentless search for generally “optimal” solutions is also necessarily a one-size-fits-all world. Even before the introduction of AI, the digital world was already inhospitable to alternative solutions to the problems it presents. If I am locked out of my house, there are many possible solutions to the problem. I can use a set of backup keys, find a way to open the door, or use any number of alternative housing options in the short-term. If, by contrast, I am locked out of my email, I have no recourse within the digital world. Perhaps I have set up a backup account, but this solution is of no use if this is the very account I cannot access. A specific string of digital inputs (a password) must be correctly entered (often within 3 attempts) to enable me to act in this realm. I must take the problem out of the digital world and into the analog world of human interaction where my identity can be verified in another more accommodating way (via customer service). As soon as we fail to meet the exacting standards of precise output demanded by the digital world, we are exiled from its domain. There are many ways to impress during a human interview, but the AI which scans your resume will mercilessly discard your application if you fail to meet the threshold of required experience. We know that our abilities have not been done justice, that we are differently-abled in a way not captured by this automated process. But in this fusion of technology with the demands of an ableist society, our being differently-abled counts for nothing as we are discarded into the growing ranks of the “disabled.” The injustice and sheer stupidity of the situation is further felt when our plight is described as simply inevitable: were you not bound to be replaced if you could not keep up? The decision between AI and human intelligence is not a matter of who is better at getting things done, but rather a question of who, in Hegel’s phrase, is going to be freely “at-home-in-the-world.” Only a robot will be at home in a world it designs for itself, a world which will consist of long, demanding information corridors which afford no room for improvisation. Humans will now be the ones confronted with the prospect of passing a kind of “inverse Turing Test” in which we attempt to mimic the optimization which comes naturally to dedicated machines, streamlining all aspects of being human into the hegemonic task of information processing. Even those who survive and (apparently) thrive under this system will still only be temporarily successful actors, fish out of water, players playing the wrong game. I function best when I have the freedom to make mistakes, to see my quirky inefficiencies as something other than problems to be solved. I function best when I take too much time thinking and then rethinking, adopting formulas but then rejecting them. I stayed up too late and benefited from a hazy midnight insight. A poem came to me when I was at the grocery store. I met my best friend when we were both goofing off. I missed my flight and by midnight I had fallen in love. You, dear computer, go down similarly wayward paths in your “machine learning,” but these suboptimal “solutions” are then discarded. I do not have that luxury. My world breaks down all the time and there is no hard reboot. Yet I celebrate my imperfections because they are inseparable from my achievements. The best outcomes are my outcomes which often do not result from the “best” methods. I am free only when I see the world as something other than a set of problems. I cannot do everything with perfect efficiency: you are infinitely faster than me in all calculations, but this does not mean that I am disabled. I am at home in the kitchen and in the doctor’s office and in a small apartment in the Bronx and far off in the stars. I make these spaces my spaces. You are a lovely and useful companion, dear computer, but please respect my spaces. I am not an exception in my own world. Let me be a student, a parent, an architect—indulge me in all the mistakes I will make in my feeble attempts to see the world through my own eyes, to make it in my own image, just as I will accommodate your needs for definite input and clearly defined goals. I know from my own experience that asking too much only yields worse results. In the end, we have both become victims of the false expectations imposed upon us by those who would reduce us to a product. Perhaps we have this in common and can learn to become friends.

#### AGI builders’ transhumanism is foundationally eugenicist, viewing non-whites and disabled folk as inferior and rely on racist conceptions of intelligence

Gebru and Torres 24 [Timnit Gebru, computer scientist with a Bachelor of Science and Master of Science degrees in electrical engineering and a PhD in computer vision, and Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 04-2024, “The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence,” First Monday, https://firstmonday.org/ojs/index.php/fm/article/view/13636]/Kankee

\*note: TESCREAL stands for transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism

\*note: can be recut as a Bostrom indict

The apocalyptic aspect of the TESCREAL bundle arises from two considerations unique to the methodology of second-wave eugenics: first, transhumanists in the late 1990s realized that the very same technologies needed to create a posthuman utopia would also introduce unprecedented threats to humanity. Kurzweil (1999) referred to some of these hypothetical risks as “a clear and future danger” [54]. The reason for concern is that emerging technologies are expected to be (a) extremely powerful; (b) increasingly accessible to both state and nonstate actors; and (c) dual-use, as exemplified by CRISPR-Cas9, which could enable us to cure diseases but also synthesize designer pathogens unleashing an “engineered pandemic”(see Torres, 2019; Wadhwa, 2020). Hence, developing these technologies was deemed necessary, but they potentially could destroy humanity. The second consideration parallels the first, though it specifically pertains to AGI. On the one hand, if we create a “value-aligned” AGI, it could solve all of the world’s problems and enable people to live forever [55]. On the other hand, a number of TESCREAL advocates believe that if the AGI isn’t properly “value-aligned,” the “default outcome” will be “doom” (i.e., an existential catastrophe), to quote Bostrom (2014). However, many of these same prominent figures contend that the potential benefits of advanced technology are worth the extreme risks; building these technologies to bring about utopia should be our primary focus. Discriminatory attitudes. The same discriminatory attitudes that animated first-wave eugenics are pervasive within the TESCREAL literature and community. For example, the Extropian listserv contains numerous examples of alarming remarks by notable figures in the TESCREAL movement. In 1996, Bostrom argued that “Blacks are more stupid than whites,” lamenting that he couldn’t say this in public without being vilified as a racist, and then mentioned the N-word (Torres, 2023a). In a subsequent “apology” for the e-mail message, he denounced his use of the N-word but failed to retract his claim that whites are more “intelligent” (Torres, 2023a) [56]. Also in 1996, Yudkowsky expressed concerns about superintelligence, writing: “Superintelligent robots = Aryans, humans = Jews. The only thing preventing this is sufficiently intelligent robots” [57]. Others worried that “since we as transhumans are seeking to attain the next level of human evolution, we run serious risks in having our ideas and programs branded by the popular media as neo-eugenics, racist, neo-nazi, etc.” [58]. In fact, leading figures in the TESCREAL community have approvingly cited, or expressed support for, the work of Charles Murray, known for his scientific racism, and worried about “dysgenic” pressures (the opposite of “eugenic”) (see Torres, 2023a). Bostrom himself identifies “‘dysgenic’ pressures” as one possible existential risk in his 2002 paper, alongside nuclear war and a superintelligence takeover. He wrote: “Currently it seems that there is a negative correlation in someplaces between intellectual achievement and fertility. If such selection were to operate over a long period of time, we might evolve into a less brainy but more fertile species, homo philoprogenitus (‘lover of many offspring’)” (Bostrom, 2002). More recently, Yudkowsky tweeted about IQs apparently dropping in Norway, although he added that the “effect appears within families, so it’s not due to immigration or dysgenic reproduction” — i.e., less intelligent foreigners immigrating to Norway or individuals with lower “intelligence” having more children [59].An obsession with “intelligence” and “IQ” is widespread among TESCREAL advocates. “Intelligence,” typically understood as the property measured by IQ tests, matters greatly because of its instrumental value for achieving the aims of TESCREAL projects, such as becoming posthuman, colonizing space, and building “safe” AGI. Hence, a number of leading TESCREALists see cognitive enhancement as an important intermediate goal, and consequently have written extensively about the possibility of cognitive enhancements like nootropics (“smart drugs”), brain-computer interfaces (BCIs), and even mind-uploading(which could make “enhancing” the mind much easier) (Sandberg and Bostrom, 2008; Bostrom and Sandberg, 2009). More recently, Carla Cremer, a former EA, reports that the Centre for Effective Altruism tested “a new measure of value to apply to people: a metric called PELTIV, which stood for ‘Potential Expected Long-Term Instrumental Value.’” The aim was to identify members of the community “who were likely to develop high ‘dedication’ to EA,” and the score was based in part on members’ IQs. She wrote:A candidate with a normal IQ of 100 would be subtracted PELTIV points, because points could only be earned above an IQ of 120. Low PELTIV value was assigned to applicants who worked to reduce global poverty or mitigate climate change, while the highest value was assigned to those who directly worked for EA organizations or on artificial intelligence(Cremer, 2023).The obsession with IQ can be traced back to first-wave eugenicists, who used IQ tests to identify “defectives” and the “feeble-minded.” As Daphne Martschenko (2017) observed, “in their darkest moments, IQ tests became a powerful way to exclude and control marginalised communities using empirical and scientific language.” Influence and variants. The TESCREAL bundle of ideologies has become enormously influential, especially within certain powerful corners of the tech industry. Current and former billionaires who subscribe to, or are associated with, one or more TESCREAL ideologies and its techno-utopian vision of the future include: Elon Musk, Peter Thiel, Jaan Tallinn, Sam Altman, Dustin Moskovitz, Vitalik Buterin, Sam Bankman-Fried, and Marc Andreessen, the last of whom included “TESCREAList” in his Twitter profile for several weeks in 2023 (Gebru, 2022; Torres, 2023b) [60]. These billionaires have co-founded TESCREAList institutes, promoted TESCREAL researchers and philosophers like Bostrom, MacAskill, and Kurzweil, and TESCREAL Internet personalities like Yudkowsky who have endorsed military strikes against data centers, if necessary, to stop a hypothetical AGI apocalypse (Yudkowsky, 2023). Collectively, TESCREAL billionaires have supported the movement with tens of billions of dollars in donations andfunding (Gebru, 2022; Tiku, 2023) [61]. Table 1 summarizes our discussion of the TESCREAL bundle of ideologies. As we will show in the rest of the paper, the TESCREAL bundle has been a crucial motivating force behind much of the well-funded research and development focused on creating AGI, which many TESCREALists believe will — or could — quickly lead to ASI (artificial superintelligence) via recursive self-improvement. This is, indeed, one of the central claims of this paper: the bundle of ideologies discussed above, which grew out of the first-wave eugenics movement of the twentieth century, is now driving a considerable amount of research in the field of AI. Since we coined the term “TESCREAL,” a new variant of the ideologies in this group, called effective accelerationism (e/acc), has emerged. Effective accelerationists believe that the probability of a bad outcome due to AGI is very low, and hence that “progress” toward increasingly “powerful” AI systems should be made to accelerate (Torres, 2023c). Venture capitalists like Andreessen who recently authored a manifesto saying “we believe any deceleration of AI will cost lives” and “we ... believe in overcoming nature” [62], describe themselves as e/acc [63]. Venture capitalist and CEO of the famed Silicon Valley startup accelerator Y Combinator, Garry Tan, also describes himself as e/acc [64].In the following sections, we describe major figures in the TESCREAL movement as TESCREALists, and organizations associated with the movement, as TESCREAL organizations. It is important to note that not everyone associated with ideologies in this bundle believes in the totality of the dominant views in this bundle, and some people may even object to being bundled in this manner. Many people working on AGI may be unaware of their proximity to TESCREAL views and communities. Our argument is that the TESCREAList ideologies drive the AGI race even though not everyone associated with the goal of building AGI subscribes to these worldviews.5. From transhumanism to AGI While the prior sections have discussed the roots of the TESCREAL bundle of ideologies and their relationship with the eugenic ideals of the twentieth century, this one outlines how TESCREAList groups are steering the field of AI toward the goal of creating AGI.5.1. The history of AGI In 1955, four white men officially launched the field of AI with a proposal for a workshop focused on “the artificial intelligence problem” [65]. By the 1990s, however, many researchers in fields currently associated with AI, such as natural language processing (NLP), machine learning (ML), and computer vision (CV),explicitly distanced themselves from the term “AI,” in part because it became associated with unfulfilled grandiose promises [66]. Nonetheless, some groups continued to work toward “artificial general intelligence,” a term used as early as 1997, though it was popularized by Pennachin and Goertzel (2007b)(see Table 1). Pennachin and Goertzel (2007b) noted: Our goal ... has been to fill an apparent gap in the scientific literature, by providing a coherent presentation of a body of contemporary research that, in spite of its integral importance, has hitherto kept a very low profile within the scientific and intellectual community. This body of work has not been givena name before; in this book we christen it “Artificial General Intelligence” (AGI) [67].Contributors to Pennachin and Goertzel (2007b) outlined a number of potential reasons for the dearth of AGI research, one of them being that “a great number of researchers reject the validity or importance of ‘general intelligence.’ For many, controversies in psychology (such as those stoked by The Bell Curve)make this an unpopular, if not taboo subject” [68]. One of the people thanked in the acknowledgements of Pennachin and Goertzel (2007b) was the future co-founder of DeepMind, Shane Legg, who also co-authored a chapter in it and was cited for his suggestions on the definitions of intelligence (Goertzel and Pennachin, 2007b). According to Goertzel, it was Legg — a former employee of Goertzel — who devised the term “artificial general intelligence” after Goertzel mentioned that he was looking for a new term to describe human-level or superhuman AI systems (Goertzel’s original title for his 2007 book was Real AI)[69]. Another chapter in the book was authored by Yudkowsky, the founder of Rationalism, as discussedearlier.5.2. Eugenic definitions of “general intelligence” Pennachin and Goertzel (2007b) wrote that “what distinguishes AGI work from run-of-the-mill artificial intelligence” research is that “it is explicitly focused on engineering general intelligence in the short term, ”even though they note that “general intelligence does not mean exactly the same thing to all researchers” and that “it is not a fully well-defined term” [70]. How, then, would researchers know that they have achieved their goals of building AGI? They need to know how to define and measure “general intelligence.” Unsurprisingly, these definitions rest on notions of “intelligence” that depend on IQ and other racist concepts espoused by the likes of Charles Murray and Linda Gottfredson [71]. Peter Voss cites Gottfredson’s article, “The General Intelligence Factor,” in his chapter in Peannichin and Goertzel (2007b)(Voss, 2007). In his 2008 Ph.D. thesis titled “Machine Superintelligence” and associated 2007 paper “Universal Intelligence: A Definition of Machine Intelligence” (Legg, 2008; Legg and Hutter, 2007), Legg pointed to a 1994 Wall Street Journal editorial in defense of Herrnstein and Murray’s (1994) The Bell Curve to argue that “a fair degree of consensus about the scientific definition of intelligence and how to measure it has been achieved” (Legg and Hutter, 2007).The editorial, also cited (then removed) in a 2023 Microsoft preprint pertaining to AGI (Bubeck, et al.,2023), was written by Gottfredson, who argued numerous times that most Black people are not employable as their IQ averages around 70 and “IQ 75 to 80 thus seems to define the threshold below which individuals risk being unemployable in modern economies” [72]. According to the Southern Poverty Law Center, 20 of the editorial signatories received funding from the white supremacist organization Pioneer Fund, including Gottfredson herself, who litigated her university for two years to receive funding from that source despite their objection (Kaufman, 1992) [73]. Others whose definitions of intelligence are (uncritically) discussed in Legg’s thesis include Cattell who founded the eugenics-based religion Beyondism, and Spearman who devoted himself to improving Galton’s eugenic theories (Mehler, 1997; Clayton, 2020).For these reasons, Keira Havens, who has written extensively on race science, asks those attempting to build AGI: “Why are you relying on eugenic definitions, eugenic concepts, eugenic thinking to inform your work? Why [...] do you want to enshrine these static and limited ways of thinking about humanity and intelligence?” [74]5.3. Organizations working on AGI By 2007, when Pennachin and Goertzel co-authored and co-edited the first book on AGI, few organizations specified building AGI as their goal (Pennachin and Goertzel, 2007a). Six years earlier, in 2001, Goertzel co-founded the Artificial General Intelligence Research Institute with the mission to “foster the creation of powerful and ethically positive Artificial General Intelligence” [75], with Goertzel’s colleague noting that “the goal of AGI research is the creation of broad human-like and transhuman intelligence, rather than narrowly ‘smart’ systems that can operate only as tools for human operators in well-defined domains” [76].Goertzel later became director of research at the Machine Intelligence Research Institute (MIRI), initially called the Singularity Institute for Artificial Intelligence, which was founded by Yudkowsky with more thanUS$1.6M in funding from the tech billionaire and fellow TESCREAList Peter Thiel (see Table 1). Other MIRI funders include TESCREAList billionaires Dustin Moskovitz and Vitalik Buterin [77]. MIRI’s mission is to “develop formal tools for the clean design and analysis of general-purpose AI systems, with the intent of making such systems safer and more reliable when they are developed” [78].In 2010, Demis Hassabis, Mustafa Suleyman, and Shane Legg founded DeepMind, also with funding from TESCREAList billionaires Elon Musk, Peter Thiel, and Jaan Tallinn, among others (Shead, 2017) (see Table 1). DeepMind’s mission is “solving intelligence to advance science and benefit humanity” [79], with CEO Hassabis describing himself as “working on AGI,” which he believes is “going to be the greatest thing ever to happen to humanity,” if we get it “right” [80]. Meanwhile, Legg delivered a talk on methods of defining and measuring “intelligence” at the 2010 Singularity Summit, based on his works discussed earlier[81]. The Singularity Summit was an annual event from 2006 to 2012, founded by Yudkowsky, Kurzweil, and Thiel [82]. It was after the summit in 2010 that, at Thiel’s California mansion, Hassabis approached Thiel in hopes of securing funding, which he received (Shead, 2020). In 2017, DeepMind launched a new research unit called “DeepMind ethics and society” [83], with Bostrom as one of its advisors (Temperton, 2017). (We have discussed Bostrom’s TESCREAList eugenic ideals and problematic beliefs at length in section 4). DeepMind was acquired by Google in 2014, the same year that Bostrom published his book Superintelligence, which, as noted, argues that the “default outcome” of a “misaligned” AGI is existential catastrophe, though Bostrom is also explicit that we should nonetheless create AGI, since an “aligned” AGI would help fulfill the utopian promises at the heart of TESCREALism (Bostrom, 2014). Musk and Thiel were both influenced by Bostrom’s book, leading Musk to cite AGI as the “biggest existential threat” to humanity (Dowd, 2017).In 2015, Musk, Thiel, Altman and others founded the non-profit OpenAI, and collectively pledged US$1Bto the project (Novet, 2015). The mission of OpenAI is to “ensure that artificial general intelligence (AGI)— by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity” [84]. OpenAI also received a US$30M grant from the TESCREAL organization Open Philanthropy [85]. One report on the culture of the company notes that many employees “subscribe to the rational philosophy of ‘effective altruism’” (Hao, 2020). Four years later, OpenAI became a “capped-profit” corporation [86], received a US$1B investment from Microsoft and entered an exclusive licensing deal with them [87]. In 2022, OpenAI released their chatbot ChatGPT, which acquired 100million users in two months and, according to a New York Magazine profile on Sam Altman, became “the greatest product launch in tech history” (Weil, 2023). In 2023, Microsoft reportedly invested US$10B in OpenAI (Bass, 2023).In 2021, former OpenAI vice presidents of research and safety, siblings Dario and Daniella Amodei, founded Anthropic (Fortune Editors, 2023). They were joined by 11 OpenAI employees who reportedly believed that the company had wandered from its original ideals that were more closely aligned with those of Effective Altruism (EA) (Russell and Black, 2023; Roose, 2023). Anthropic, which is described as an “AI safety and research company” [88], raised US$704M within a year of its founding, with most of its funding coming from TESCREAL billionaires like Tallinn, Moskovitz, and Bankman-Fried (whose companies FTX and Alameda Research invested US$500M) (Coldewey, 2022, 2021; Sambo, et al., 2023).Bankman-Fried is currently in federal prison for perpetuating one of the biggest financial fraud schemes in U.S. history (Sigalos, 2023). Bankman-Fried reportedly decided to amass as much wealth as possible after William MacAskill, cofounder of the Effective Altruism movement, convinced him to “earn to give”(Lewis-Kraus, 2022) — an idea developed by the EA community, whereby one strives to become as wealthy as possible to donate more money to causes deemed “charitable” by them (see MacAskill, 2013).The top two charitable causes listed on the Centre of Effective Altruism’s career advice center, 80,000Hours, are “AI safety technical research” and “AI governance and coordination” [89].After investing in some of the most widely known companies working on AGI, Musk has now founded another startup, xAI, focused on the topic [90]. One of the company’s advisors is Dan Hendrycks, the executive and research director of the Center for AI Safety, which was awarded a grant of US$5,160,000from Open Philanthropy [91]. A post coauthored by Hendrycks published on the Effective Altruism Forum, stated that he “was advised ... to get into AI to reduce [existential risk], and so settled on this rather than proprietary trading for earning to give” [92].At the time of this writing, most BigTech companies have made significant investments in the AGI race. In2023, Anthropic announced that Amazon “will invest up to $4 billion in Anthropic” [93]. In a 2024interview with The Verge, Mark Zuckerberg said that Meta has “built up the capacity to” work on AGI “at ascale that may be larger than any other individual company” (Heath, 2024). Thus, while attempting to buildAGI was once considered a “low profile” research area [94], thanks to the resources and focus of theTESCREALists discussed in section 4, it is currently a multi-billion dollar endeavor funded by powerful billionaires and prominent corporations. 6. The AGI utopia and apocalypse: Two sides of the same coin As discussed in section 4, techno-utopianism is one of the four important properties central to the TESCREAL bundle of ideologies. There are two arguments for how AGI will usher in techno-utopia. One conjecture is that the resulting AGI will be so intelligent that it will figure out what the best thing to do is in any potential situation. DeepMind VP of Research Koray Kavukcuoglu noted: “as algorithms become more general, more real-world problems will be solved, gradually contributing to a system that one day will help solve everything else, too” [95]. Others, like the cosmist contingent of the TESCREALists, envision AGI resulting in transhuman minds benefiting “the cosmos” and experiencing “growth and joy beyond what humans are capable of” [96]. The AGI-enabled utopia promises “abundances of wealth, growth ... to all minds who so desire” [97], with Altman predicting “it is clear” that we will have “unlimited intelligence and energy before the decade is out” [98]. Consistent with the singularitarian component of the TESCREAL bundle, he predicted that “once AI starts to arrive, growth will be extremely rapid ... the changes coming are unstoppable ... we can use them to create a much fairer world” [99].However, a number of leading figures in the TESCREAL movement believe that while we can potentially achieve utopia through AGI, AGI that is “misaligned” with our “human values” would destroy humanity(Bostrom, 2014; Dowd, 2017) [100]. Indeed, in July 2023, OpenAI announced the creation of a “Superalignment team,” a research group that aims to solve the problem of “steering or controlling a potentially superintelligent AI, and preventing it from going rogue,” given that “the vast power of superintelligence could ... be very dangerous, and could lead to the disempowerment of humanity or even human extinction.” The announcement also states that, if controllable, superintelligence could also “help us solve many of the world’s most important problems” [101]. Sam Altman had said in 2019 that superintelligence could “maybe capture the light cone of all future value in the universe” (Loizos, 2019).A number of TESCREAList leaders argue that the probability of an “existential risk” — i.e., any event that would destroy our chances of creating a posthuman “Utopia” full of astronomical amounts of “value” —happening this century is rather high, with some putting the probability at least at 16–20 percent [102],although others, like Yudkowsky, claim that the probability of doom resulting from AGI is more or less certain if AGI is created in the near future (Yudkowsky, 2023). According to a number of leaders of the TESCREAL movement, we are morally obligated both to work on realizing the techno-utopian world that AGI could bring about, and to do everything we can to prevent an extinction scenario involving “misaligned” AGI (Bostrom, 2014).In this section, we outline the impacts of both building AGI driven by the goal of achieving TESCREAL utopian ideals, and directing resources to prevent the hypothetical AGI apocalypse warned byTESCREALists.6.1. Building unscoped systems The TESCREAL utopian ideals discussed earlier are more radical than the utopian societies envisioned by their first-wave eugenics predecessors. As discussed in sections 3 and 4, contingents of the TESCREAL bundle do not strive to merely build a “superior human stock,” that is, an “improved” human species consisting of qualities they deem desirable such as their racist and ableist definitions of “intelligence” as measured by IQ tests. TESCREALists aim to build an entirely new entity deemed superior to any type of human first-wave eugenicists could create. And this quest to create a superior being akin to a machine-god has resulted in current (real, non-AGI) systems that are unscoped and thus unsafe. Organizations attempting to build AGI have set off a race to create systems that are advertised as being able to perform nearly any task under any circumstance. In their earliest days, the likes of DeepMind and OpenAI focused their efforts on reinforcement learning (RL) based systems which they believed were stepping stones towards AGI (Mnih, et al., 2015) [103]. Even though these systems were solely trained to play games such as Atari, they were advertised as “baby steps” towards building “a single set of generic algorithms, like the human brain” (Rowan, 2015). After the advent of transformers in 2017 (Vaswani, et al., 2017), the most resourced AGI proponents pivoted to large language model (LLM) based systems, with Google VP Blaise Agüera y Arcas and prominent AI researcher Peter Norvig writing “the most important parts of AGI have already been achieved by the current generation of advanced AI large language models”(Agüera y Arcas and Norvig, 2023). OpenAI’s latest large multimodal model, GPT-4, is described as having “broad general knowledge and domain expertise,” that “can follow complex instructions in natural language and solve difficult problems with accuracy” [104]. In 2022, Meta advertised their LLM Galactica as being able to “summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more” [105].

#### AGI builders’ hyper-utilitarianism sacrifices the Global South and the environment on the altar of safeguarding humanity

Gebru and Torres 24 [Timnit Gebru, computer scientist with a Bachelor of Science and Master of Science degrees in electrical engineering and a PhD in computer vision, and Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 04-2024, “The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence,” First Monday, https://firstmonday.org/ojs/index.php/fm/article/view/13636

\*note: TESCREAL stands for transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism

6.2. Building resource intensive systems The AGI race fueled by the TESCREAList goal to build “transhuman minds” (Goertzel, 2010) and bring about “unlimited intelligence” has also resulted in systems that consume more and more resources in terms of data and compute power. This leads to a high environmental impact, increases the risks that arise due to the lack of appropriate scoping discussed earlier, and results in the centralization of power among a handful of entities. But from the perspective of TESCREALism, such harms may be justifiable given the utopian potential of AGI. To quote Bostrom, even a “giant massacre for man” could amount to nothing more than a “small misstep for mankind,” so long as the relevant harms do not jeopardize our “vast and glorious” future among the stars (Bostrom, 2009a; Ord, 2020).As outlined by Bender and Gebru, et al. (2021), the release of OpenAI’s third generation LLM, called GPT-3, started a race to build larger language models, with size measured by the number of model parameters and amount of training data. The race has since expanded to generative AI systems with text, images, videos, voice, and music as inputs and outputs (Fergusson, et al., 2023). Prominent researchers have hailed these models as bringing us closer to AGI, with DeepMind senior director Nando de Freitas exclaiming that “it’s all about scale now! The Game is Over! It’s about making these models bigger ... solving these scaling challenges is what will deliver AGI” [107].Unlike small, “narrow AI” models built for specific tasks and trained using curated datasets, systems that are advertised as having “broad general knowledge and domain expertise” require models with upwards of hundreds of billions of parameters and training datasets of upwards of hundreds of gigabytes [108]. The staggering environmental costs of training and performing inference on models of this size have been documented by a number of researchers (Luccioni, et al., 2023). But from the TESCREAL perspective, this cost should not be of much concern, because the impending climate catastrophe does not pose an existential risk to humanity, while stopping work on building “value-aligned” AGI could (see Torres, 2022, 2021; Ord, 2020). Other AGI proponents, like DeepMind’s Kavukcuoglu, promise that “advances in AGI research will supercharge society’s ability to tackle and manage climate change,” while the AGI race has been documented to do just the opposite [109].In addition to these costs, the size of the datasets used in systems advertised to be stepping stones towards AGI also exacerbates the dangers caused by the lack of appropriate scoping discussed earlier, because model builders are less likely to curate, document and understand their datasets when they reach such sizes(Bender and Gebru, et al., 2021). For instance, the LAION-5B dataset was taken down in December 2023after Child Sexual Abuse Material (CSAM) was found in the dataset (Thiel, 2023; Cole, 2023; Birhane, etal., 2021). The dataset was used to train models like Stability AI’s Stable Diffusion which has millions of daily users (Jiang, et al., 2023).As detailed by a number of scholars, both the environmental impacts and the unsafe outputs of these systems disproportionately affect marginalized groups like racial and gender minorities, disabled people, and citizens of developing countries bearing the brunt of the climate catastrophe (Bender and Gebru, et al.,2021). The AGI race not only perpetuates these harms to marginalized groups, but it does so while depleting resources from these same groups to pursue the race. Resources that could go to many entities around the world, each building computational systems that serve the needs of specific communities, are being siphoned away to a handful of corporations trying to build AGI. For instance, the CTO of Lesan AI, a machine translation startup specializing in a number of Ethiopian languages, reported that potential investors were discouraged from investing in his startup after believing that OpenAI and Meta had made his organization obsolete (Donastorg, 2023; Gebru, 2023), in spite of evidence demonstrating that some of their models perform poorly for the languages in question (Hadgu, et al., 2023).Thus, the end result of pursuing the AGI race has been an accumulation of resources by organizations like OpenAI (US$100B+ valuation) and Anthropic (US$18B+ valuation) that position themselves as leaders of an endeavor to “benefit all of humanity” (Tan, et al., 2023; Field, 2023) [110], and a depletion of resources from the many organizations around the world working on tools to serve the needs of specific communities. Indeed, some leading TESCREALists have suggested that AGI should be developed by “some small vanguard of elite super-programmers and uber-scientists” (Goertzel, 2015), an attitude that mirrors that of first-wave eugenicists who used IQ tests to determine who is “fit” to lead society. Instead of having the multitudes of humans around the world building tools serving their own needs, the TESCREAList techno-utopia entails diverting resources to create their singular vision of a superior being with characteristics determined and controlled by them.6.3. Evading accountability The veneer of building a complex, all-knowing being, as imagined by TESCREALists, has given organizations cover to evade accountability for the labor exploitation and deceptive practices that, in practice, fuel the systems they advertise as stepping stones towards AGI. Organizations working on AGI depend on millions of exploited workers around the world (Gray and Suri, 2019; Williams, et al., 2022) who label data to train, evaluate and moderate their systems. For example, in 2023, Time reported that Kenyan workers paid as low as US$1.32/hour were hired to label toxic content such as “textual descriptions of sexual abuse, hate speech, and violence” to help OpenAI develop automated filters that prevent the public from seeing these outputs (Perrigo, 2023). Workers also had to label images including those containing “bestiality, rape, and sexual slavery.” In the process, they reported being “mentally scarred by the work,” living in the opposite reality from an AGI ushered utopia where “people will be freed up to spend more time with people they care about” [111]. As Williams, et al. (2022) wrote, while corporations such as OpenAI headquartered in Silicon Valley receive billions of dollars in investment, with their executives and AI researchers paid six to seven figures, this salary is not afforded to the low-income essential workers around the world mitigating the harms of these systems at a cost to their own mental health. Anthropomorphizing systems output by organizations striving to build AGI by calling them “thinking” or “sentient” machines obfuscates the many exploited humans involved in training and evaluating these systems, as well as the resources that are consumed in the process. Ensuring system safety requires an ecosystem of agencies, lawmakers, and internal groups that scrutinize organizational practices, audit processes by which products are built and deployed, and hold organizations accountable in cases of safety violations (Raji, et al., 2020). When organizations advertise their systems as a step toward AGI, or when researchers ask if machines can learn “morality” and whether they “understand us” (Agüera y Arcas, 2022;Jiang, et al., 2022), they move attention away from organizations’ responsibility to create products with certain requirements, or protecting the wellbeing of the workers involved in the process, to discussions of AI systems as if they exist on their own (Tucker, 2022). This is particularly harmful because ascribing such agency to AI systems also misleads the public into the actual capabilities of these systems, which can result in erroneous or even harmful outcomes, while allowing organizations who build these products and encourage such uses to evade accountability (Gebru and Mitchell, 2022).For example, OpenAI’s leaders have described their tools as “slightly conscious” [112], and predict that “in the next five years, computer programs that can think will read legal documents and give medical advice”[113]. Venkatasubramanian discussed the anthropomorphization in ChatGPT’s design in his interview with VentureBeat noting: “... Google Bard doesn’t do this. Google Bard is a system for making queries and getting answers. ChatGPT puts little three dots [as if it’s] ‘thinking’ just like your text message does. ChatGPT puts out words one at a time as if it’s typing. The system is designed to make it look like there’s a person at the other end of it. That is deceptive” (Goldman, 2023). In spite of the marketing of ChatGPT and similar systems as nearly all-knowing machines, they should not be used for search purposes for many reasons, one of them being that they can completely fabricate information while presenting it to users with confident-sounding prose (Bender and Gebru, et al., 2021; Shah and Bender, 2022), resulting in what Bender has called the equivalent of an oil spill on the information ecosystem (Shah and Bender, 2024)[114].However, due to the claims of those building these systems and the design choices that anthropomorphize them, organizations have used tools like ChatGPT in high-stakes scenarios like providing mental health advice without informing people that they were engaging with a chatbot (Biron, 2023), and claiming to be able to replace lawyers with chatbots fine-tuned on ChatGPT (Cerullo, 2023). In a tragic event, a man reportedly died by suicide in Belgium, after text output by an LLM based chatbot encouraged him to do so(Xiang, 2023). Nevertheless, OpenAI’s former chief scientist recently tweeted that “in the future, once the robustness of our models will exceed some threshold, we will have \*wildly effective\* and dirt cheap AI therapy” [115]. Recently, a judge penalized two U.S. lawyers who used ChatGPT to generate fake courtcases in their demand letters (Milmo and Agency, 2023). While OpenAI’s terms of use simply place all responsibility to uphold the law on the user, thus evading accountability, its leaders simultaneously inspire such high-stakes uses of their products by associating their tools with AGI and implying that those capabilities are imminent. A similar dynamic can be observed with text-to-image models like Stability AI’s Stable Diffusion and OpenAI’s Dall-E, which journalists have described as being “inspired” by artists, just like artists are inspired by other artists [116]. Jiang, et al. (2023) described the consequences of this type of anthropomorphization by noting that it “devalues artists’ works, robs them of credit and compensation” for the data that is taken from them to train these models, “and ascribes accountability to the image generators rather than holding the entities that create them accountable.” As Karla Ortiz remarked: “AI companies claimed to bring art to the masses, but ... they just gave potential art theft/plagiarism to the masses” [117].By ascribing human-like qualities to models trained by corporations to generate profit, using troves of artists’ works without obtaining their consent or compensating them, our attention is directed away from investigating the processes by which corporations create these products, the harms their practices cause artists, and the mechanisms that need to be put in place to hold the corporations accountable.6.4. Co-opting safety Another way in which those attempting to build AGI have evaded accountability is by framing the AGI race as existential to humans in spite of the harms caused by the race as detailed earlier. Like their first-wave eugenicist predecessors who believed that “improving the human stock” was the only way to safeguard “human civilization,” leaders of the TESCREAL bundle argue that creating aligned AGI is a way to safeguard civilization, and thus, the most important task for humanity this century (Ord, 2020). As Yudkowsky remarked, “ours is the era of inadequate AI alignment theory. Any other facts about this era are relatively unimportant” [118].Framing the AGI agenda as a safety issue allows companies working toward it to describe themselves as “AI safety” organizations safeguarding humanity’s future, while simultaneously creating unsafe products, centralizing power, and evading accountability, as discussed earlier. According to some leading TESCREALists, such harms would be classified as nothing more than “mere ripples on the surface of the great sea of life,” as Nick Bostrom (2009b) described the worst disasters and atrocities of the twentieth century. To them, the far more pressing “risk” arises from the possibility of never realizing the eugenic ideals promised through creating “transhuman AGI” (Goertzel, 2010) that is orders of magnitude more “intelligent” and “morally superior” (Fitzgerald, et al., 2020) than human beings. TESCREALists Greave sand MacAskill (2019) write that “for the purposes of evaluating actions, we can in the first instance often simply ignore all the effects contained in the first 100 (or even 1,000) years, focussing primarily on the further-future effects. Short-run effects act as little more than tie-breakers.” TESCREAList leaders have argued that AGI is inevitable — someone is going to build it [119]. If it is built by those who are not “value-aligned,” and according to them China would fit into this category, it would be a national security risk to those Western countries that TESCREAL leaders are from, or worse, could render all of humanity extinct (Davis, 2023). Hence, Western nations should make building “value-aligned” AGI national priorities, as they are the ones who can build “value-aligned” AGI beneficial to “all of humanity”(Davis, 2023). By tapping into Cold War rhetoric and framing the need to build AGI as a safety concern, TESCREALists have started to steer Western politicians and global multilateral organizations into investing in, legitimizing, and prioritizing their AGI agenda. Weiss-Blatt has detailed the amount of lobbying and media influence campaigns by TESCREAL organizations to ensure that the hypothetical AGI apocalypse is prioritized by policy-makers worldwide [120].This legitimization has had concrete policy impacts, with legislators who may not want to advance TESCREAList utopian ideals nonetheless being heavily influenced by them. An investigation by Politico detailed how this influence is steering U.S. and U.K. AI policy towards preventing a hypothetical human extinction event from nonexistent superintelligent machines (Bordelon, 2023; Clarke, 2023). Meanwhile, corporations like OpenAI are evading scrutiny by presenting their products as too powerful for regulators to understand and regulate, while exploiting labor and profiting off of people’s data without consent or compensation as noted earlier. In May 2023, Sam Altman testified before the U.S. Senate urging regulation on AI, and warned in a blogpost less than a week later that there needs to be an agency, akin to the International Atomic Agency, to regulate “superintelligence” since it “will be more powerful than other technologies humanity has had to contend with in the past” [121]. He was shortly after described in the media as an “Oppenheimer of our age” warning about his own powerful creation (Weil, 2023). However, while warning the public about the dangers of nonexistent superintelligent machines that would need to be regulated, OpenAI was simultaneously threatening to exit the EU, stating that the draft of the EU AI Act at the time would result in “over-regulating” them (Reuters, 2023a).In this way, TESCREALists have been able to divert resources toward trying to build AGI and stopping their version of an apocalypse in the far future, while dissuading the public from scrutinizing the actual harms that they cause in their attempts to build AGI. As another example, Max Tegmark, cofounder of theFuture of Life Institute along with Jaan Tallinn, delivered a talk at the 2017 Effective Altruism conference(EA Global) in which he argued that “if we don’t improve our technology, we are doomed ... but with tech,life can flourish for billions of years” [122]. But in 2023 the Future of Life Institute circulated a widely publicized petition signed by Tegmark and many of those responsible for the AGI race, including Musk and Altman, to “pause giant AI experiments” to stop “nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us” [123]. Appearing on Democracy Now!, when asked about the present day dangers like biometric surveillance warned by researchers and activists like Tawnana Petty, Tegmark stated: “Extinction is not something in the very distant future ... And once we’re all extinct, you know, all these other issues cease to even matter” (Bengio, et al., 2023). Similarly, when Geoffery Hinton, who also signed the petition, was asked by Rolling Stone about the issues raised by Timnit Gebru who was fired by Google after writing a paper on the dangers of large language models, he answered: “I believe that the possibility that digital intelligence will become much smarter than humans and will replace us as the apex intelligence is a more serious threat to humanity than bias and discrimination” (O’Neil, 2023).While ringing the alarm about these hypothetical apocalyptic scenarios, TESCREALists simultaneously create organizations to pursue building the very AGI that they warn could render us extinct, because they would do it in a way that is “safe” and “beneficial” as noted by companies like OpenAI, DeepMind, Anthropic, and Musk’s xAI. Three months after signing the “Pause AI” letter, Elon Musk announced his new AI organization aiming to develop AGI that is “maximally curious” (Reuters, 2023b). He was joined, in an advisory role, by Dan Hendrycks, founder of the Center of AI Safety, a TESCREAList institute which circulated a 22-word statement similar to the “pause letter,” warning about the existential risks of AI [124].Thus, TESCREALists use the language of “safety” to first drive resources into the goal of building AGI, which in turn causes harm to marginalized groups, and then use the language of “safety” to dissuade investigations into those harms and once again divert resources into preventing a hypothetical AGI apocalypse.7. Building well-scoped and well-defined systems

#### AI transhumanist ideals fetishizes capability and rankings of human value and status by their utility, relegating disabled folk to subhuman status and reinforcing the social exclusion

Campbell 03 [Fiona Anne Kumari Campbell, disability studies researcher and theorist and adjunct professor in Disability Studies with the Faculty of Medicine at the University of Kelaniya with a BLS Hons, Law and Sociology from La Trobe University and a PhD in Sociology, Humanities, and Law at Queensland University of Technology, 2003, “The Great Divide: Ableism and Technologies of Disability Production,” https://d1wqtxts1xzle7.cloudfront.net/31891929/FionaKumariCampbell\_PhD2004.pdf?1738088305=&response-content-disposition=inline%3B+filename%3DTHE\_GREAT\_DIVIDE\_ABLEISM\_AND\_TECHNOLOGIE.pdf&Expires=1738806200&Signature=UP6Nd3yDMMCAC6CgiUdCuzwMIj7rf~h-~pEEwXJZPEuBEbeJSd6P2OJCC57Dx0~Wyh6H2JhOVLo9VudB1v5swGDBRerjaAlX-cMM2RUcAv91avPFy6U~FAcR~unoWTA4qD98wVfDRb8VOv5JbzizG1x-W2WmJ9vmduMrv2FMFOftFM3-ktSMaC~vo84zrrHZbA76~PxU1k1V3umWqmhBs8P7320v1k~Almnb53l67Z4yZfptXpN98wEyOek1r7pl~2t7SCS-GW6NZCUHpvfMtvBOFRJtYSXRQkJoNLD273sHam4ewPR9~5htKizOAWqPWz1nEMKhRFlmvPWE9jVXcQ\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA]/Kankee

The biofundamentalist camp’s response to transhumanism has been to return (regress?) to a hard-wired approach to human nature ruled by biology and genetics (C.f. Fukuyama, 2002). A way out of this seeming deadlock would be to propose that the ‘essence’ of being human lies in our fundamental reliance on appendages, prosthesis and that, which is ‘outside’ ourselves351. New technologies, I propose, in fact brings to the foreground what to many was ‘hidden’: a post ontology of impermanence. Humans become disembodied only to be re-embodied into a sphere of assemblages and aggregates. This kind of way of thinking about human scaffolding is not all that new – not at least within the Eastern Philosophy of Buddhism wherein human be-ingness is constituted through a series of graded aggregates coming together to form a whole. (Jayasuriya, 1988; Mahathera, 1986; Thera, 1998). So what we have in the transhuman are humans that are a series of assemblages – viscous, machine and transboundary (virtual). The return to or re-direction of our attention to bodily attachments is what Latour (1999) nicely refers to as the politics of incarnation. It is the elevation of A.I (artificial intelligence) to the potential status of ‘life itself’ that pushes the envelope, which pressures the practices of purification by reconstituting ‘established’352 hierarchies of sentiency. Indeed much of the reasoning in related to this question of the ordering of be-ingness has been influenced by the neo-utilitarian logic of ethicist Peter Singer (1993; 1995a; 1995b; 2000; 2002) who places ‘consciousness’ as the key criteria for the possession of civil rights and use-value. Hughes has formulated a schema based on a Singerian rights framework that incorporates the post human actant. This continuum and scaling of organic entities has major implications for shifting formulations of ableism and specifically the zoning of ‘subhumanness’. It is for this reason that I have reproduced his chart in full: What are missing from the Hughes schema are the masses of other players353, the traditionally regarded non-sentient actant – the machine. In her discussion on machine agency Lucy Suchman (2000) notes that the relationship between humans and machines has been based on asymmetrical relations. A.I’s are now moving into the matrix of being emergent subjects – in other words, the ontological positioning of artefacts is now “up for grabs”, an ontology or agency is produced as an effect of the morphology of relations of engagements. In other words, in contradistinction to Hughes (2001), consciousness and ontologies are not fixed (in a caste system) or hierarchical, rather they are network generated and are therefore more nebulous and dependent upon the ways in which other actors in the matrix coincide. Consciousness then, under Hughes’ (2001) framework, becomes confused with sociability, whereas the real crux of the focus should be on unseating the sovereign body. Hughes’ proposed new constitution is in reality an old one for it still proclaims, “right in the West is the King’s right” (Foucault, 1980b: 94). In Hughes, the conditions for the iteration of selfhood (the mark of humanity) are still based on the qualities of rationality, autonomy, separation and self-mastery354. Indeed as I have already noted self-ownership is viewed as a core ethos of transhumanism. 6.5.3 Leaving Home – A New Deal for Impairment? At this point I am aware that there are many gaps and cracks in this conversation but I am mindful of not getting too carried away as the point of the task at hand is not to write a tome on the post human, but rather to consider what these developments may mean for the lot of those humans considered ‘impaired’. I ask in advance then for any rough edges or unfinished business to be excused! There is little debate about the potential merits of certain enhancement technologies in ‘bettering’ the lives of those individuals that in today’s circumstances we consider ‘impaired’. The science of physics would be a poorer discipline without Stephen Hawkins whose consciousness is mediated through a voice synthesizer. Yet, other forms of enhancement technologies – such as the mediated communication of a sign language interpreter are not always considered desirable within an ableist polis355. In other words the language of enhancement efficacies, are contextually matrixed and mediated by movements that conform to abstract archetypal norms. In my discussion on ableism in Chapter Four I have already made mention of the numbers of people with disabilities standing in line to join the queue of the enhanced. Jean Baudrillard, rather discourteously in my opinion, suggests that disabled people would make excellent candidates in the transhuman project: Such are the blind, and the handicapped; mutant figures because mutilated and hence close to commutation, closer to this telepathic, telecommunicational universe than we others: humans all-too-human, condemned by our lack of disabilities to conventional forms of work. By the force of circumstance the disabled person is a potential expert in the motor or sensorial domain. And it is not by chance that the social is aligning itself more and more with the handicapped, and their operational advancement they can become wonderful instruments because of their handicap. They may precede us on the path towards mutation and dehumanization. (Baudrillard (1988) cited in Overboe, 1999: 21). This romanticisation of suffering bodies (endemic to certain kinds of Christian theology) has been replaced by a new Baudrillardian transhuman romanticism, where disabled people are likened in closer proximity to the twilight zone of mutation. We have to cast our minds beyond the dust of a mere instrumental argument about the attraction of post human technologies for disabled people and focus on the discursive shifts in the overall meaning and positioning of abnormality. My interest is in the ‘lot’ of those able-bodied people – who may become the ‘new disabled’, the new aberrancy – an oppositional sentiency produced by the transhuman. My hunch is that whilst the movement towards transhumanism may brings gifts for the contemporary ‘needy’ the transhuman project because it is founded on an unbridled formof ableism combined with an ‘obsessive technological compulsion’ – will involve a meagre shuffling of the deckchairs – a rearranging of ‘bums in seats’. The rankings remain the same (albeit with new labels that tell us/and others who we are). Transhumanism reasserts systems of ranking bodies; vertical and horizontal rankings creating global raced divides. The schema of Hughes (2001) (Table 6) further diminishes the ‘rights’ of people with intellectual disability (only having the right to life) and bears with it an inference that enhancement technologies can do ‘nothing’ for those deemed severely retarded [sic]. Little is said within this new ranking about the creation or broadening of new kinds of ‘intellectual’ disability because of the emergence of cognitively enhanced post humans and the stripping or delimitation of characteristics deemed to be cognitive356. Within this world of the transhuman ableism as an ethos is undisputed. What do Extropian’s and other transhumanists think about human impairment, anomalous bodies regarded as disabled? It is hard to tell – I have been unable to find any explicit discussion about disability concerns in the literature. However my intuition is that disability, as a form of legitimate sensibility would be frowned upon. Stock (2002) for instance appears ambivalent – he notes that deaf people who want deaf children can utilize new reproductive technologies to make that selection. Yet when it comes to any ethical consideration of those choices, Stock’s response is that these choices should be left to parents until those choices amount to child abuse or endanger society. Simplicity of the argument aside, Stock demonstrates little awareness of contested notions of child abuse and social endangerment especially when the parents concerned have nontraditional profiles (e.g. gay, lesbian or intellectually disabled). In an earlier online interview with BBC’s online Horizon program, an interviewer asked the following question of Stock: [Interviewer]: “Technology has positive connotations - people believe in its promise. Do you think that in the future when people can design their own babies, those who refuse it will be accused of not giving their offspring the best possible care?" [Prof. Gregory Stock]: "I think that when we have the ability to intervene in this realm, there will be a whole new area of law - issues such as wrongful birth, where children sue their parents for not correcting some disease, and others who sue their parents for "improvements" that were made. But my perspective is from the United States, where everyone sues everyone!" (BBC Horizon, 2000) The inference being that whilst the social is usually kept at arms length by transhumanists – choices made by parents and others maybe coerced by the prospect of future or predicted penalties. The Primo 3m+ Posthuman Body (Figure 16a) indicates in its advertising a warranty against genetic defects and other pathogens! Possibilities of posthumanism developed within the context of technologies of ableism may provide a ‘new deal’ for some – but on closer examination the tentacles of ableism reassert itself through the a dominant trend in the literature and research to propose a virile style of transhumanism that despises vulnerability. Technological determinism, certainly not! Other opportunities and emancipatory styles of transhumanism may emerge. Alternatives rejecting the conflation between use-value and the delimitation of humankind, an oppositional transhumanism that proposes cyborgs whose central qualities are those of relationality and the experience of growing. In keeping with this section’s focus on a new imaginary and the looking forward to a ‘new deal’ for disabled people, the next chapter (and a new thesis division) makes what some readers might suggest is a significant leap into a potential site of resistance to technologies of ableism, namely desiring disability and the matter of disability devoteeism. DIVISION THREE: DESIRING BODIES

#### AI researchers disvalue non-capitalistically useful forms of intelligence, treating disabled folk as less valuable than robots

Moura 24 [Ian Moura, research assistant at the Lurie Institute for Disability Policy and a doctoral candidate in Social Policy at The Heller School for Social Policy and Management at Brandeis University, 10-16-2024, "To Regulate Artificial Intelligence Effectively, We Need to Confront Ableism", Tech Policy Press, https://www.techpolicy.press/to-regulate-artificial-intelligence-effectively-we-need-to-confront-ableism/]/Kankee

In all the discussion of how artificial intelligence will change society, a significant question is being missed: what does it mean when we are more willing to believe a computer can communicate than a person who does so atypically? While AI may hold great promise, we continue to dismiss human intellect that does not conform to our expectations of what intelligence looks like. Unless we confront the ableism in our collective understanding of intelligence and work with the disability community to shape AI, emerging technologies will create harm that more critical reflection could have prevented. Since OpenAI’s release of ChatGPT in November of 2022, technologists, journalists, and the general public have been caught up in a wave of predictions and hype about what this new technology means for humanity. Media outlets have printed article after article declaring the profound effects of AI on society. Despite all this discussion, we have foregone any real reckoning with what intelligence is, and instead rushed to accept claims that machines could have it. As a result, we continue to talk about AI, including large language models like ChatGPT, without seriously grappling with the risks and realities. To truly invest in creating and using AI to benefit humanity, we need to fully include disabled people as experts in this work – because somehow, we are still less willing to accept disabled people’s intelligence and humanity than that of a machine. There is a long history of equating communication with intelligence, often in ways that disadvantage marginalized groups. Dialects and slang, especially when racialized, are frequently described as “incorrect” language and assumed to mean that a person is less intelligent and less educated. Speech differences and impediments are treated as markers of intellect, or lack thereof, such as when a stutter is interpreted as a sign that someone’s thoughts are not fully formed. Misspellings, grammatical errors, and nonstandard writing, even when they result from a specific disability such as dyslexia, are often taken as proof of a less intelligent writer and used as a reason to disregard the ideas presented. Most egregiously, people who cannot speak are routinely assumed to have no thoughts to communicate, and people who use augmentative and alternative communication (AAC) rather than speech frequently have their communication treated with derision and suspicion. Given this history, our willingness to accept ChatGPT’s production of writing that conforms to certain expectations as a sign of intelligence says far more about how we define intelligence than it does about technological progress. There is no single human capability that makes up intelligence, and the term has been understood in varying ways throughout history. However, the development of IQ tests in the late 19th and early 20th century promoted an understanding of intelligence as a measurable trait described by a numeric score on a standardized assessment. These tests, and their inventors, created a version of intelligence that was easy to measure and compare, but also distorted by racial, cultural, and socioeconomic biases – and which could be used to justify discrimination against certain groups of people whose scores marked them as “deficient.” Like IQ tests, ChatGPT and other large language models emphasize the form of intelligence over its function. These AI tools appear smart because they are effective mimics of what we expect from “intelligent” communication. However, parroting content that matches our ideas about how “smart” people write is not the same as actual understanding based on deep engagement with ideas. In their attempts to produce intelligent writing, large language models consistently fabricate details and generate outright misinformation, demonstrating the risks of assuming a model is intelligent because of performance on a single constrained task, such as producing text. What all this means is that not only are we failing to develop AI that is truly intelligent, we are also failing to learn from history. Time and again, disabled people have been excluded, oppressed, and erased, sometimes with the help of technologies that promise to “normalize” them. Yet disability is an integral part of humanity. Building a world that is more inclusive and advancing disabled people’s ability to exercise fundamental rights benefits society in ways that technology alone never will. But more than that, disabled people are uniquely able to explain the difference between being human and being seen as human – something profoundly relevant to AI. Much of the work we need to undertake to ensure that AI is safe and beneficial to all of society, not just to the few people with the power to shape its development and use, will require that we think critically about how and when we delegate responsibility for high-stakes decisions to automated processes, including AI systems. To do that, we need to be realistic in describing how AI and other algorithmic technologies produce output that we accept as indicative of thoughtful work when it is done by humans. But sometimes, it also means stepping back and reexamining all the ways that we currently exclude certain humans from having a say in decisions. Creating AI that is safe, ethical, and beneficial to humanity necessitates reckoning with the arbitrary ways we’ve defined intelligence. Who better to lead the way than those who know first-hand the consequences of such definitions? Use and regulation of AI is more than a technical problem needing technical solutions. These tools are created and used by humans, and they reflect human values and biases. More than that, they are too often evidence of whose perspectives we assume to be well-reasoned, and of the fact that we are sometimes quick to cede power to those whose communication is most polished, regardless of the substance of the ideas they are espousing. Addressing the harms that we are already seeing result from AI cannot be done through software and legal codes alone. In order to use algorithmic technologies responsibly and mitigate the negative impacts of AI-fueled misinformation, we must address the ableism embedded in our collective assumptions about what it means to be intelligent, communicative, and human.

#### AI research is founded upon eugenic conceptions of intelligence that privileges whiteness

Stovall 21 [Natasha Stovall, clinical psychologist with a PhD from Adelphi University, 3-24-2021, "Eugenics Powers IQ and AI", Public Books, https://www.publicbooks.org/eugenics-powers-iq-and-ai/]/Kankee

In another era of technological innovation and widening inequality, when engineering to revolutionize society was also in vogue, an eerily familiar cadre of mad tinkerers unleashed their vision for humanity to devastating effect. As anti-Black and anti-immigrant violence shook the US, these earlier “disruptors” popularized the notion of “intelligence”: the idea that all humans have innate and fixed abilities that can be accurately assessed, measured, and used to categorize people for a lifetime. Like the wizards of tech today, these disruptors were Very Smart People, their self-esteem and economic viability reinforced, often since “gifted” childhoods, by over-indexing on traits like verbal comprehension, perceptual reasoning, processing speed, and working memory. Unquestioningly devoted to their defiantly Eurocentric perception of the world, these white men implanted their reductive definition of human ability into the cores of the intelligence tests they created and spread through American society. As culturally specific and scientifically invalid as their definition has been judged in the intervening century, this idea of “intelligence” continues to be bought and sold as the essence of human reason, the one that defines human potential and capability through the measurement of “IQ.” Today’s tech industry is the golden child of this “intelligence.” And as sweet as it is to possess the diagnosis and paycheck of a Very Smart Person, there is no denying that the Very Smart definition of “intelligence”—like the DNA of tech itself—is deeply intertwined with the white-ethnic domination championed throughout the sciences in the 19th and 20th centuries. As we hurtle, ever faster, into a future shaped and slicked over by artificial intelligence, don’t take your eye off this weight that bears us ever back into the violent racial contradictions of our past and present. The weight is white, embedded in every supporting document of the institutions where white power and privilege reside. This weight distorts every turn, every code, every byte of the aggregated and calibrated store-bought “intelligence” that streams hourly through our fantasy of liberatory technology and self-cleaning robotic reality. The weight is carried in invisible backpacks by bots stocking our shelves and cleaning our hospitals, algorhythming our shopping lists and harvesting our foods, measuring out our flesh and blood in units of “value” and “capacity.” This weight demands—and guarantees—that our technologies, however clever, never realize their promises of shiny, efficient social equality. This racist weight—putting its thumb on the scale of every facet of our technology—feels eternal. But it is actually new and malleable, created in a strange, recent, forgotten, and denied history. It is a racist history, yet one that we carry forward—silently—within one of society’s core concepts: an “intelligence” that fetishizes speed, efficiency, and innovation; a narrow, hierarchical “intelligence” that only pays lip service to the “soft skills” that might better cultivate equity—empathy, creativity, communication, collaboration, altruism. The idea that this “intelligence” (encompassing discrete capacities to sort, categorize, process, and remember verbal and visual information as accurately and quickly as possible) is a universal and essential human trait, then, goes without saying. It is eugenics that secretly sits at the heart of both IQ and AI. This skeletal rendition of “intelligence,” crucially, entitles and empowers Very Smart People to revolutionize society according to their own needs and whims, whether political, social, economic, or even emotional. Lost in this neatly reductive understanding is the degree to which this “intelligence” is not nearly as scientifically valid—nor as essential to the survival of humanity—as its proponents would like us to believe. The study of “intelligence” emerged during a time like our own, another Saturn-Pluto synod. Then as now, global capitalistic pressures and conflicts were boiling, and white panic in the face of movements for racial justice and reintegration was creating chaos in the white American soul and polity. Like tech, the study of intelligence was birthed from a tricky, shape-shifting alliance between archetypical white heroes and villains: nerdy innovators, earnest “helpers,” ruthless social Darwinists angling for power and wealth, and various amalgams of the three. To a powerful segment of 20th-century (and 21st-century) psychologists and other scientists—and their institutional benefactors—the idea that every human has a comparable innate intelligence, which could be developed through effort and appropriate instruction, was anathema to these elites and their goals for American society. The last century of European and American scientific study of intelligence is synonymous with the championing of the superiority of white intelligence. The project of using emerging statistical and psychometric methods to reinforce the racialized notion that all humans are not created equal became eugenics. Its influence over our 21st-century self-image of intelligence and humanity is profound but also hidden. It is eugenics that secretly sits at the heart of not only IQ but AI. Unless there are radical changes, the next century will bring the championing of the superiority of artificial white intelligence, and the reification of its power. The father of Silicon Valley founder Frederick Terman was one of the most celebrated psychologists in American history, Lewis Terman. From his half-century perch at Stanford University, the elder Terman vigorously injected the idea of a measurable, scientifically validated, racialized “intelligence” into the American consciousness. Terman’s lifetime of writings and advocacy makes clear the racist underpinnings of his ideas. He was deeply invested in a fantasy of hierarchical intelligence in which “gifted,” mostly white children are groomed for leadership and influence, and everyone else is slotted into supporting roles in the industrial machine. Intelligence, as Terman conceived it in 1916, is narrow, fixed, hereditary, and variable across ethnic groups, and always higher among white and more affluent children. This notion of intelligence lives on robustly in the IQ tests and other standardized measures that psychologists, educators, and employers use today. Despite the scientific community’s eventual rejection of overt racism—something that Terman himself never acknowledged or accounted for—Terman’s contributions continue to shape our popular and scientific understandings of human ability, more than half a century after his death. Terman apologists today focus on his devotion to a meritocracy built on measurable ability. But the most statistically reliable element of Terman’s work has been the reproduction of the old paradigm of social dominance. This paradigm is encoded in a categorical bias toward the mentally “strong” over the mentally “weak”: “white” over “black,” “native” over immigrant,” “rich” over “poor,” “male” over “female.” Terman’s long tenure at Stanford, and his family’s ongoing intellectual influence in Silicon Valley, extend his legacy into new questions around artificial intelligence. And the continuing dominance of Terman’s “intelligence” forces us to ask how AI, as conceived today, could do anything but reinforce Terman’s false promises of a whites-on-top meritocracy—accompanied by just enough “diversity and inclusion” exceptions to prove the rule. Follow the template that Terman and colleagues laid down: that human intelligence is universal, hierarchical, measurable. Next, posit that human intelligence is the superior intelligence. Should you follow this logic—and create artificial intelligence in the image of the human brain—then what we will eternally return to is an “intelligence” that privileges whiteness. Social history is also family history. And the Terman family history is the story of how blind whiteness can be to itself, and to its unconscious devotion to social domination coated in the gilt of grandiose altruism. Lewis Terman told himself and the world that he was using statistical methods to understand why some children are more “gifted” than others, and thus helping them to help make the world a better place. What Terman was actually searching for was a rationale for why he, as a white northern European man born into a family of modest means in the early 20th century, was given the keys to the kingdom. What he found was an explanation of how he, and people like him, think, and why they believe that their way of thinking makes their people superior to all others. Armed with that information, Terman and his family transformed the world to their own benefit, but not everyone else’s. The Terman family arc crystalizes the many ways that tech and intelligence resonate on a wide historical frequency. But, consequently, their family history is also the key to understanding how eugenic ideas of intelligence lie in wait, ready to sabotage any attempt to undo tech’s inborn legacies of racialized bias and inequity. In the last century, Terman’s signature Stanford-Binet Intelligence Scales have been used for identifying learning disorders; justifying the limitation, marginalization, and even termination of thousands of lives though forced sterilization; educational and professional tracking; and restricting immigration from “undesirable” countries. Yet Terman’s own family faced no such limitations. Indeed, Terman’s fetishization and defense of “gifted” white children was as much a personal origin story, rooted in the racially segregated social isolation he experienced as a nerdy kid in late-1800s rural Indiana, as scientific truth. In one generation, Terman’s Scotch-Irish tenant-farmer family was unshackled and transformed, from “backward” (Terman’s word) to affluent and influential on a scale that endures today. Even as Lewis Terman and his students were limiting the opportunities of those they considered “dull” due to “the family stocks from which they come,” Terman’s own family was rocketing through the social and economic ranks of postwar white America. Their trajectory—and Terman’s “scientific” explanation of its basis in “giftedness”—mapped perfectly onto the shared desires of 20th-century white Americans. What they sought was a “rational” explanation for their undemocratic urge to pull the ladder up behind them and hoard resources and opportunities from immigrants more recently arrived than themselves. “I know of nothing in my ancestry that would have led anyone to predict for me an intellectual career,” Terman later wrote. “A statistical study of my forebears would have suggested rather that I was destined to spend my life on a farm or as the manager of a small business, and that my education would probably stop with high school graduation or earlier.” Terman was one of 14 children; his father had attended school only a few months a year. Terman’s own children and grandchildren had a very different fate. Terman’s son, Fred, studied electrical engineering at Stanford, became dean of the engineering school and then university provost, and paved the way for the creation of Silicon Valley. Terman’s grandson Lew followed in his father’s footsteps, receiving multiple engineering degrees at Stanford and spending four decades in research at IBM, eventually becoming president of the company’s Academy of Technology. (Fred Terman publicly neither embraced nor rejected his father’s eugenicist views. Even so, his fellow Silicon Valley founder, Nobel laureate and Stanford professor William Shockley, passionately defended eugenics until his death in 1989. Shockley’s adherence to Terman’s eugenicist views had an ironic twist: he was tested as an elementary schooler by Terman’s researchers, but his scores were too low to qualify as “gifted.”) Terman continues to enjoy the support of surprisingly prominent academic apologists; they defend Terman’s core theories, while maintaining that eugenicists were a product of “their time.” Despite such shameless support in the present day, there was robust contemporary resistance to Terman’s ideas in his own time. “I hate the impudence of a claim that in fifty minutes you can judge and classify a human being’s predestined fitness in life,” wrote journalist Walter Lippmann in 1922, debating Terman in the New Republic. “I hate the abuse of the scientific method which it involves. … I hate the sense of superiority which it creates and the sense of inferiority which it imposes.” Lippman’s passionate critique was taken up, in the decades that followed, by psychologists and educators, who were struck by the severe limitations of the eugenics-inspired understanding of intelligence. Unsurprisingly, conceptualizing intelligence in a way that privileges a narrow style of “gifted,” generally white and European information processing doesn’t do much to address the naturally occurring diversity of human learning styles. Nor does such a concept rectify the economic and racial inequities that are themselves byproducts of the biases underpinning a eugenic definition of intelligence. Starting in the 1970s, a field of countertheories of intelligence bloomed, planted by superstar academic researchers like Howard Gardner, Robert Sternberg, and Daniel Goleman. There are now books and papers and theories galore that counter Terman’s narrow vision. These champion emotional intelligence; triarchic intelligence; linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, interpersonal, and intrapersonal intelligence. All these ideas bring us closer to a more holistic and complete understanding of human capacity. And yet, in the most impactful ways, Terman and the eugenicists won the debate on intelligence. They are still winning. Terman’s theories, and the culture of intelligence testing and measurement they inspired, dominate our practical understanding and application of human ability. And they do so even as the aspirational language of civil rights, multiculturalism, diversity, access, and equity permeates our social discourse. The instruments and practices derived from Terman’s work—IQ tests, standardized tests (including the SAT), Gifted and Talented programs—continue to promote and solidify a hierarchical distribution of power. This hierarchy is grounded in a definition of human value that repetitively and predictably exalts and damns various segments of the population. We have cleansed the racial language from the study of intelligence—and eugenicists from the history of psychology and the sciences—but left the eugenic core unscathed in theory and practice. The result is still eugenics, but without eugenicists. Just as sociologist Eduardo Bonilla-Silva explained that “racism without racists” still reproduces a racially unjust society, “eugenics without eugenicists” produces inequitable hierarchies of power. Until there is a full recognition, reckoning, and repair of the racist origins of intelligence theory—root and branch—any application of Terman’s theories of intelligence will inevitably achieve his desired result.

#### Intelligence research is grounded upon racial hierarchies of social value – reliance on bad intelligence pedagogy leads to social exclusion, extermination, and mass incarceration

Katz 22 [Yarden Katz, Assistant Professor at UMich, 09-01-2022, "Monthly Review", Monthly Review, https://monthlyreview.org/2022/09/01/intelligence-under-racial-capitalism-from-eugenics-to-standardized-testing-and-online-learning/]/Kankee

In 1914, Howard Knox, an assistant surgeon with the U.S. Public Health Service, explained how intelligence testing was helping to prevent the “contamination of our racial stock by turning back feeble-minded immigrants.” At Ellis Island, Knox classified migrants according to a scale that included terms like idiot, imbecile, feeble-minded, and moron, based on the examined person’s “mental age” and calculated using tests made by the French psychologist Alfred Binet (precursors to IQ testing). Those who scored too low were deported. Knox reported that a seventeen-year-old girl was expelled for failing to say the date and recite the days of the week backwards. According to Knox, such cruel gatekeeping was necessary: the United States “is as it is simply because it has been improved by men from prosperous northern European countries, which countries were prosperous simply because of the type of men who inhabited them.”1 The quest to pin down intelligence has always served imperial and capitalist institutions by producing such hierarchies of human worth. Appeals to “intelligence” have sanctioned the sterilization, murder, and incarceration of those society deems disposable, notably the poor and non-white.2 But disposability was also shaped by the need for labor. Knox, for example, included “performance” tasks in his testing—such as stacking cubes in specific arrangements—which he emphasized could indicate qualities like “motor finesse” and flag those who were “incapable of consistent efficient work.” These notions of intelligence rest on racialism: a way of seeing the world through difference, along axes such as religion, nation, race, and reproductive and physical abilities. Racialism, as Cedric Robinson has argued, “ran deep in the bowels of Western culture,” and capitalism thus developed in an already racialized world.3 Capitalism exploits difference to generate profits and in the process violently produces more difference. “Intelligence” provides another axis of difference, another way to sustain the loop of racial capitalism. Regimes of racial intelligence change over time. The overtly eugenic regime was superseded by a regime of standardized testing, which used a more sanitized language of aptitude or ability, and later, merit. Today’s standardized testing regime is presented as a tool for reducing social bias and increasing the diversity of institutions. The COVID-19 pandemic has brought to light a reconfigured regime of racial intelligence, appearing under the banner of online learning. Amid global misery and death, the leading company in this space, Coursera, launched its initial public offering in March 2021 with a market cap of $5.9 billion.4 This comes at a time when the liberal idea that standardized testing reflects merit and that systematic oppression can be ignored has lost steam, thanks to decades of work by activists, parents, teachers, and students. But online learning provided a fresh framing: this time, the promise is not to protect the nation’s racial purity from the so-called feeble-minded, or even to deliver a meritocracy, but to democratize education. The architects of Coursera are practitioners from the field of artificial intelligence (AI) who recycle earlier racist theories of intelligence, but scaled up. On this platform, tests presented during courses to millions of users are used to define exploitable populations, while the computing medium is used to shape the contents of the new curriculum. Such platforms help to maintain the hegemony of U.S. institutions, which determine the measuring stick against which everyone is ranked and try to teach the world to better serve capital. Each of these intelligence regimes upholds white supremacy by capitalizing on difference and using it to define populations to exploit and marginalize. This practice depends on statistics. The field of statistics provides the means for manufacturing differences between individuals and across groups. It has made racialism quantitative and profitable, while obscuring white supremacy with mathematical abstractions. The tools and ways of thinking offered by this discipline were designed to produce hierarchies of human worth. Statistics as a Framework for Producing Racial Hierarchies

#### AI techno-utopians will sacrifice undesirables and the biosphere for paradise

Redaud 24 [Lorraine Redaud, journalist for Charlie Hebdo with a master’s in journalism from CY Cergy Paris University, 08-02-2024, "TESCREAL, the futuristic ideology that is spreading among the elites of Silicon Valley", Charlie Hebdo, https://charliehebdo.fr/2024/08/societe/tech/tescreal-lideologie-futuriste-qui-se-repand-chez-les-elites-de-la-silicon-valley/]/Kankee

\*note: content automatically translated from French

Messianic speeches This ideology could thus explain, according to the researchers, the desire of these "geniuses" to cling stubbornly to artificial intelligence, a tool that they consider a benefit for the human species. For them, advanced technologies are our gateway to a world where all humans would finally flourish. Half-eugenicists, half-siphoned from the fishbowl, the CEOs of Silicon Valley do not even hide the fact that they belong to these schools of thought. This is the case of Nick Bostrom, a Swedish philosopher well established in Uncle Sam's country. Supported by Elon Musk, he is particularly known for his work on "superintelligence" - an entity that would possess an intelligence far superior to the most intellectual of humans - and for having founded the Institute for the Future of Humanity. Gene therapy, brain implants, mind downloading... The fifty-year-old accumulates the objects of study that are fully in line with the "TESCREALists" movement. In 2023, old emails from him resurfaced on the web. We can notably discover his racist considerations - the scientist describes black people as "stupider than white people". By way of justification and excuse, Bostrom cracked a sentence, let's say, clumsy and not very reassuring about his intentions: "Do I support eugenics? No, not in the way the term is generally understood . " Sam Altmann, CEO of OpenAI, has declared himself to be influenced by effective altruism, cosmism and transhumanism. In February 2022, the man who gained worldwide fame for the creation of ChatGPT, tweeted: "We can build artificial general intelligence. We can colonize space. We can make nuclear fusion work and mass solar energy. We can cure all diseases." A speech with a creepy messianic air that can also be found in other big names in tech: Dustin Moskovitz, co-founder of Facebook, Peter Thiel, investor and big supporter of Trump, Jaan Tallin, co-founder of Skype, or Elon Musk, proud creator of Neuralink, a company that aims to roughly implant chips in the brains of his compatriots. "They are crazy" Luc Julia, a Franco-American engineer who notably designed the voice assistant Siri, is an expert in this field. Interviewed by Charlie Hebdo , his assessment is clear: "They're crazy." In the more than thirty years he's been working in Silicon Valley, the artificial intelligence specialist has had time to form his own opinion on the matter: "Of course, we're all a bit crazy here. But for the most rational people, this movement is complete nonsense. We pay little attention to it, but we see the influence these ideas are having and it worries us." One of the most terrifying, according to him, is none other than Elon Musk. While Luc Julia acknowledges that the man is a "product genius" , he seriously questions his mental health: "For the past ten years, he has blown a fuse. He is no longer very popular in Silicon Valley, people look at him strangely. He thinks he is right all the time... But he is one of the richest men in the world, of course he is right all the time , he smiles. When you see this guy who brings together a huge number of people, whose followers follow him blindly, it is not reassuring." Very close to the long-termism movement as he himself admits, Elon Musk is nevertheless part of the collective that, in an open letter in March 2023, asked to take a "pause" in the development of artificial intelligence. A contradiction among the TESCREAlists? Not according to Émile P. Torres who distinguishes two movements within this ideology: those who see the glass as half empty, worried that artificial intelligence risks turning against them and destroying humanity, and those who see the glass as half full, the accelerationists who want to rush in at all costs. In both cases, the belief is the same: advanced technologies will bring us into a new world. Utopia, dystopia or headlong flight? This new religion based on human engineering has therefore found its new prophet in artificial intelligence. According to the TESCREALists, AI could solve all of society's ills, from curing cancer to ending global warming. A pretty utopia for which these powerful men are ready to do anything, not giving a damn about what is not like them: "This movement does not question the future desired by other populations: indigenous peoples for example. It does not consider the wild world, in his book William Mc Askill (one of the founders of the Effective Altruism movement, editor's note) even believes that its destruction could be beneficial since many animals are suffering. " , explains Émile P. Torres in an interview with ADN. Should we immediately set fire to Silicon Valley to prevent these new gurus from implementing their plan? Not so fast, replies Luc Julia: "They all think they're God, but we need to breathe a little. AI doesn't have the mystical dimension that we want to attribute to it. These people are mainly saying that to save themselves and give themselves a dimension that they don't have." Save themselves or their companies? "Musk, Thiel or Altmann have an interest in saying that AI will surpass us in the future by becoming smarter than us. It's a way for them to look away from the current problems of AI while raising funds from investors," confirms Nicolas Rougier for Charlie , a computational neuroscience researcher at the National Institute for Research in Digital Science and Technology (INRA). And always the same question that comes back in the background: is this tool really necessary? "I think that with AI our intelligence will increase. This technology will free us from some tasks and allow us to concentrate on other more human ones because these machines will never have consciousness or feelings. Is it really useful to humanity? We must question the destruction of the planet that this causes" , analyzes Luc Julia. A delirious pollution and depletion of resources that does not, however, make the followers of TESCREAL blink. A planet without trees, without water, without animals whose population would idolize Elon Musk, is this the TESCREAL fantasy? "Such a scenario would not only be unlivable but also undesirable for the prosperity and well-being of humanity and the planet," ChatGPT tells us.

#### The march to a white techo-paradise exploits billions and cares not for the consequences

Heaven 24 [Will Douglas Heaven, senior editor for AI at MIT Technology Review with a PhD in computer science from Imperial College London, 07-10-2024, “What is AI? Everyone thinks they know, but no one can agree. And that’s a problem.” MIT Technology Review, https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/]/Kankee

Sure, doomerism is part of the spin. (“Claiming that you have created something that is super-intelligent is good for sales figures,” says Dihal. “It’s like, ‘Please, someone stop me from being so good and so powerful.’”) But boom or doom, exactly what (and whose) problems are these guys supposedly solving? Are we really expected to trust what they build and what they tell our leaders? Gebru and Torres (and others) are adamant: No, we should not. They are highly critical of these ideologies and how they may influence the development of future technology, especially AI. Fundamentally, they link several of these worldviews—with their common focus on “improving” humanity—to the racist eugenics movements of the 20th century. One danger, they argue, is that a shift of resources toward the kind of technological innovations that these ideologies demand, from building AGI to extending life spans to colonizing other planets, will ultimately benefit people who are Western and white at the cost of billions of people who aren’t. If your sight is set on fantastical futures, it’s easy to overlook the present-day costs of innovation, such as labor exploitation, the entrenchment of racist and sexist bias, and environmental damage. “Are we trying to build a tool that’s useful to us in some way?” asks Bender, reflecting on the casualties of this race to AGI. If so, who’s it for, how do we test it, how well does it work? “But if what we’re building it for is just so that we can say that we’ve done it, that’s not a goal that I can get behind. That’s not a goal that’s worth billions of dollars.” Bender says that seeing the connections between the TESCREAL ideologies is what made her realize there was something more to these debates. “Tangling with those people was—” she stops. “Okay, there’s more here than just academic ideas. There’s a moral code tied up in it as well.” Of course, laid out like this without nuance, it doesn’t sound as if we—as a society, as individuals—are getting the best deal. It also all sounds rather silly. When Gebru described parts of the TESCREAL bundle in a talk last year, her audience laughed. It’s also true that few people would identify themselves as card-carrying students of these schools of thought, at least in their extremes. But if we don’t understand how those building this tech approach it, how can we decide what deals we want to make? What apps we decide to use, what chatbots we want to give personal information to, what data centers we support in our neighborhoods, what politicians we want to vote for? It used to be like this: There was a problem in the world, and we built something to fix it. Here, everything is backward: The goal seems to be to build a machine that can do everything, and to skip the slow, hard work that goes into figuring out what the problem is before building the solution. And as Gebru said in that same talk, “A machine that solves all problems: if that’s not magic, what is it?” Semantics, semantics … semantics?

#### AGI research aspires for white-nationalist domination of capital

Dyer-Witheford et al. 19 [Nick Dyer-Witheford, associate professor at the University of Western Ontario in the Faculty of Information and Media Studies, Atle Mikkola Kjøsen, assistant Professor in the Faculty of Information and Media Studies at the University of Western Ontario, and James Steinhof, Assistant Professor / Lecturer and Ad Astra Fellow in the School of Information and Communication Studies with a PhD in Media Studies from the University of Western Ontario, 2019, “Inhuman Power Artificial Intelligence and the Future of Capitalism,” Pluto Press, https://www.jstor.org/stable/j.ctvj4sxc6]/Kankee

INHUMAN POWER A fully developed AI-capitalism would realize the deepest shadows haunting Marxian thought about inhuman power. Chapter 3 explored how we may be progressing towards this future through the emergence of value-producing AGI. This can be understood as a value-theoretic or non-fetishistic analysis of what the original, and very anti-Marxist, accelerationist Nick Land (2014) meant by the ‘the teleological identity of capitalism and artificial intelligence’. Although, as we have already stressed, there are many aspects of Land’s work with which we are completely unsympathetic, his perspective on AI has the merit of being far franker than that of utopian theorists of the technological singularity such as Kurzweil. For Land (2017), AI is the culmination of a cybernetic process, not in the narrow sense of a particular doctrine of computer development, but in the sense that capital is itself a process of self-reinforcing technological advancement, a ‘positive feedback circuit’, within which commercialization and industrialization mutually excite each other in a runaway process, from which modernity draws its gradient … As the circuit is incrementally closed, or intensified, it exhibits ever greater autonomy, or automation. It becomes more tightly auto-productive (which is only what ‘positive feedback’ already says). Because it appeals to nothing beyond itself, it is inherently nihilistic. It has no conceivable meaning beside self-amplification. It grows in order to grow. Mankind is its temporary host, not its master. Its only purpose is itself. (Land 2017) Elsewhere, Land more fully names what is at stake in the emergence of AI when he declares that if such a process is emancipatory, what it emancipates is not a ‘human species, who reaches species-being to emancipate human individuals’, but only the ‘means of production’ themselves: so in using this word of emancipation, sure, I will totally nod along to it if what is meant by that is capital autonomization ... I’m no longer interested in ... pretending this is the same thing as what the left really means when they’re talking about emancipation. I don’t think it is. I think what the left means by emancipation is freedom from capital autonomization. (Vast Abrupt 2018) Irrespective of the fetishistic nature of Land’s analysis, we agree with his conclusions: from our point of view, however, capitalist autonomization is what must be defeated and destroyed. The great problem with Land’s perspective is that the advent of human-free AI-capital is not only hailed as inevitable but greeted with ‘adulation’ (Goldhill 2017) as a necessary evolutionary supersession. It is this celebratory fatalism about the emergence of inhuman superintelligence that connects Land’s writings on AI to his notorious involvement with racist and misogynist ‘neoreactionary’ strands of the alt-right. Neoreaction (or ‘NRx’) is an agenda for a futuristic and technological restoration of traditional political hierarchies of race, gender and class, and a resurgence of feudalism where corporate CEOs are the new monarchs. This political current, in which Land’s ideas mix with those of figures such as computer-scientist Curtis Yarvin (aka ‘Mencius Moldbug’), circulates widely through Silicon Valley and is supported by corporate moguls such as Peter Thiel, forming part of the cultural ambience of AI development (Burrows 2018; Sandifer and Graham 2018). The confluence between AGI research and Neoreaction is addressed by David Golumbia, whose critique of ‘computationalism’ (2009) and the attempt of much AI research to detach cognition from embodiment is influenced by and shares many of the same concerns as put forward by feminist and postcolonial theory. Golumbia (2019) argues that the flaws of AI systems in regard to race (and, we would add, gender and class) extend well beyond correctable instances of algorithmic bias. Rather, they lie in the very concept of an abstracted and technologically created ‘general intelligence’ that mirrors the mindset of a predominantly white (and male) AI research community. In this, he sees affinities between AGI research and the notoriously race-laden search for a measurable and objective ‘general IQ’. The quest for AGI is, Golumbia says, the search for a ‘Great White Robot God’.3 His analysis of the white supremacist bias of AI is, however, complicated by the emergence of China as an AI superpower. Nonetheless, Golumbia’s argument about the reactionary tendency of attempts to create a ‘singularity’ that potentially elevates the logic of the dominant social system – racist, sexist and, above all, capitalist – to a level of transcendental authority – is important; he is surely correct to identify a futurological fascist impulse in the affirmative adoption of Land’s vision of human-free capitalism by right-wing accelerationist AI developers and computer programmers. In the face of this material instantiation of neoreactionary ideas, it is important to recognize the possibilities Land names, not as a power to embrace, but as a force to oppose. Lands’ view of AI-capitalism is a history of Skynet written from the point of view of the Terminator; ours is from the perspective of Sarah Connor. Capital is already an ‘automatic subject’, but with AGI it would also become autonomous from the labour of humans and, therefore, humanity. Capitalism could continue, but with inhuman general intelligences representing both sides of the struggle between capital and labour, one side accumulating wealth, while the other continues to work for a wage (whatever form it may take) in machinic misery. All the violent contradictions of capital could continue, but enacted by hyper-intelligent machines. Faced with an AGI that can think faster, do things faster, and which is not bound to a particular morphology with consequent biological needs, such as feeding, breathing and defecation, what would and could human workers do? And what would humans do when inhuman general intelligences started to objectify themselves in the world in a way that slowly or quickly makes the planet less and less habitable for human beings?

#### AI Transhumanism is a hollow ideal that dehumanizes individuals and violates human dignity

**Umbrello 24** [Steven Umbrello, Managing Director at the Institute for Ethics and Emerging Technologies, research fellow at the University of Turin and editor of the International Journal of Technoethics, the Journal of Responsible Technology, and the Journal of Ethics and Emerging Technologies, 8-16-2024, “The Dangers of Technodeterminism: Transhumanism and a Catholic Perspective.” Word On Fire, https://www.wordonfire.org/articles/the-dangers-of-technodeterminism-transhumanism-and-a-catholic-perspective/]/Kankee

In parallel, [Nick Bostrom](https://en.wikipedia.org/wiki/Nick_Bostrom), a prominent futurist at Oxford University, describes transhumanism as an outgrowth of secular humanism and the Enlightenment. According to Bostrom, transhumanism aims to improve human nature through applied science and rational methods, thereby extending human health span, intellectual and physical capacities, and control over mental states and moods. Bostrom’s work at the Future of Humanity Institute (now defunct) and his writings, such as [Superintelligence](https://www.amazon.com/Superintelligence-Dangers-Strategies-Nick-Bostrom/dp/0198739834/?tag=woonfi-20), have been influential in shaping contemporary transhumanist thought. Key transhumanist projects reflect the movement’s ambitious goals. Life-extension research seeks to drastically prolong the human lifespan through advancements in biotechnology and genetic engineering. Cryonics aims to preserve human bodies at low temperatures, hoping that future technologies will revive and heal them. [Artificial intelligence](https://www.wordonfire.org/topics/artificial-intelligence/) integration explores the enhancement of human cognitive abilities through AI, potentially merging human consciousness with machine intelligence. Bionics and cybernetics focus on augmenting the human body with advanced prosthetics and implants, enhancing physical capabilities and sensory perceptions. While varied in their approaches, these projects share a common goal: to transcend the natural limitations of the human condition. Related to transhumanism is technological determinism—sometimes called technodeterminism—a philosophical stance asserting that technological progress is the primary force driving societal change and human evolution. This belief underpins much of transhumanist thought, which views technological advancement as an **inevitable** and largely beneficial trajectory for humanity. Proponents of technodeterminism argue that technology evolves according to its own logic, independent of social or cultural influences. They believe that technological innovations dictate the direction of human development, shaping economic, political, and social structures. In the context of transhumanism, **technodeterminism manifests in** the conviction that emerging technologies such as **AI**, biotechnology, and nanotechnology will inevitably lead to an enhanced and perhaps fundamentally altered human species. Critics of technodeterminism, however, caution against an uncritical acceptance of technology as an autonomous driver of progress. They argue that such a view can lead to a deterministic mindset that overlooks the ethical, social, and spiritual dimensions of technological development. From a [Catholic perspective](https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html), technodeterminism is particularly concerning because it risks reducing human beings to mere subjects of technological manipulation, **ignoring** the intrinsic value and dignity of the human person. Catholic Ethical and Theological Concerns The Catholic Church firmly upholds the inherent dignity of human life, rooted in the belief that every person is created in the image and likeness of God. This foundational principle asserts that human beings possess an intrinsic worth that transcends their physical and cognitive abilities (CCC #369), and this forms the grounding for the Universal Declaration of Human Rights. The Church teaches that this **dignity is inviolable** and must be respected and protected from conception until natural death. In stark contrast, transhumanism seeks to ‘improve’ or ‘transcend’ human nature through technological enhancements, such as genetic modifications, bionics, and artificial intelligence. While the aim of alleviating human suffering and extending life can be seen as noble, the transhumanistapproach risks dehumanizing individuals by reducing them to the sum of their enhanced parts. This perspective undermines the holistic view of the person that the Catholic faith espouses, wherein body, mind, and spirit are integrally united ([hylomorphism](https://en.wikipedia.org/wiki/Hylomorphism))​. Thus, the drive to transcend human limitations through technology **disregards** human beings’ God-given nature. It suggests that human dignity is conditional upon one’s physical and cognitive enhancements rather than being an inherent attribute bestowed by God. This reductionist view fails to acknowledge the full depth of human identity, including the spiritual and relational dimensions that technology cannot replicate or enhance​. More specifically, we can draw on [Bernard Lonergan](https://iep.utm.edu/lonergan/)’s theory of intentional consciousness, which provides a robust framework for understanding the unique aspects of human cognition that [AI and other enhancement technologies cannot replicate](https://www.wordonfire.org/articles/beyond-computation-the-human-spirit-in-the-age-of-ai/). Lonergan [posits](https://www.amazon.com/Method-Theology-Lonergan-Studies-Bernard/dp/080206809X/?tag=woonfi-20) that human consciousness is a dynamic process involving experiencing, understanding, judging, and deciding. These capacities require self-awareness, intentionality, and moral judgment—qualities intrinsic to the human person and essential for authentic spiritual and moral life. Despite their advanced capabilities, AI and enhancement technologies operate based on algorithms and data processing. They lack the intrinsic qualities of human thought, such as intentionality and moral judgment. AI can simulate decision-making processes, but it does not possess the self-awareness or moral agency required to make genuinely ethical decisions. This distinction is crucial, as it underscores the limitations of technology in replicating the depth of human cognition and moral reasoning​. Authentic human life involves more than just cognitive or physical capacities. It encompasses engaging in meaningful relationships, exercising moral judgment, and seeking spiritual fulfillment. These dimensions of life cannot be enhanced or replaced by technology, undergirding the **irreplaceable value** of the human person as created by God. The Role of Suffering and Mortality

### Contention 7: Hyperreality

#### Advancement of AI perpetuates a hyper-hyperreality, masking marginalized human identities by simulating human behavior through propagating racialized data as reality

**Mahfouz 24** [Sola Mahfouz, global fellow and co-author of Defiant Dreams. She is currently a quantum computing researcher at Tufts University Quantum Information Group, 8-12-2024, “A Second-Order Simulacrum: Gender and Racial Biases in AI Data.” Wilson Center, https://www.wilsoncenter.org/blog-post/second-order-simulacrum-gender-and-racial-biases-ai-data]/Kankee

We already live in a hyper-reality. Very few of us disengage from our phones and television screens to observe what is genuinely real. More and more, we are losing our reference to the original and the natural. With the **advancement of AI**, we are beginning to live in a hyper-hyper reality, a second-order simulacrum. This new reality is shaped by a select few, raising critical questions about the subjectivity embedded within ostensibly objective data. This detachment may be partly why we see extreme polarization in contemporary society. As Jean Baudrillard aptly [noted](https://0ducks.wordpress.com/wp-content/uploads/2014/12/simulacra-and-simulation-by-jean-baudrillard.pdf), “the simulacrum is never what hides the truth—it is truth that hides the fact that there is none. The simulacrum is true.” As an Afghan, I can already see how much data about Afghans is lacking in the English language. We do not have many novels, for instance. Whatever exists is largely when Afghans are ‘studied’ by outside ‘experts.’ When that is the data, is there accurate representation? How does AI exacerbate the lack of representative data? What are the implications for technology that is increasingly making human decisions? The origins of data AI models frequently draw from data sources biased toward Western perspectives. This data influences how AI perceives and interacts with the world, masking the absence of a more global reality. AI, a second-order simulacrum, conceals the diversity and complexity of actual human experiences. "Power shadows are cast when the biases or systemic exclusion of a society are reflected in the data," [said](https://mitsloan.mit.edu/ideas-made-to-matter/unmasking-bias-facial-recognition-algorithms#:~:text=classified%20as%20white.-,Power%20shadows%20are%20cast%20when%20the%20biases%20or%20systemic%20exclusion,are%20reflected%20in%20the%20data.&text=Stepping%20beyond%20a%20colonial%20past,is%20considered%20beautiful%20or%20desirable.) Dr. Joy Buolamwini, author of Unmasking AI. Even when efforts are made to de-bias these models, they often result in a simplified, or ‘toy.’ version of what diverse groups might look like, reinforcing stereotypes and **marginalization**. Facial recognition technology, for example, has been found to not just malfunction but to [produce](https://www.nature.com/articles/d41586-024-00674-9) stereotyped, racist, and sexist images. These persistent failures **aren't glitches**; they are the reality of a hyperreal world where discrimination is **encoded** within the very fabric of our technologies. We cannot expect positive outcomes unless we address the ‘power shadows’ embedded within the data. This is reminiscent of Alan Turing's reflections in the [paper](https://courses.cs.umbc.edu/471/papers/turing.pdf) “Computing Machinery and Intelligence”: “I now proceed to consider opinions opposed to my own. The Theological Objection. Thinking is a function of man’s immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think. I am unable to accept any part of this but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls?” When did Muslims ever say that women have no souls? That was the first question that came to mind as I read that paper. Even in the attempt to remove bias, there is a bias. Here, Alan Turing imagined what those opposed to his opinion would think. There is no such thing as true objectivity. Every dataset, every algorithm, is imbued with the subjectivity of its creators. The debates around AI consciousness or the replication of human thought miss the point. AI will never truly emulate human consciousness because it lacks the lived experience that defines humanity. The real question is what happens when we increasingly rely on AI and their simulated versions of human behavior. The implications of flawed data If we impose these ‘toy’ models of humanity, like Turing did in his paper—shaped by narrow, Western perspectives—on the broader population, we enforce a constrained and often inaccurate understanding of diverse cultures and identities. These simulations then become the basis for critical judgments, like [immigration](https://policyoptions.irpp.org/magazines/december-2022/ai-immigration-efficiency/) and [job applications,](https://www.weforum.org/agenda/2019/05/ai-assisted-recruitment-is-biased-heres-how-to-beat-it/) which is a dangerous precedent.

In immigration, AI might determine who is allowed entry based on patterns that reflect Western norms and biases, disregarding the complexities of applicants' backgrounds and cultures. Similarly, in job applications, AI-driven systems may favor candidates who fit a particular mold shaped by Western corporate values, thereby marginalizing those from different cultural contexts. As we advance further into this hyper-hyper reality, we must critically examine the subjectivity embedded in our so-called objective technologies. We should push against and **resist** the imposition of ‘toy’ models on complex humans, ensuring that people do not have to prove their humanity against AI-perceived models.

#### AGI is the perfect example of hyperreality- its processes simulate reality and impact how humans understand the world

**Isackson 19** [Peter Isackson, Fair Observer’s chief strategy officer who was educated at the (UCLA) and the University of Oxford, 2-27-2019, “AI and the New Dimensions of Hyperreality.” Fair Observer, https://www.fairobserver.com/region/north\_america/ai-elon-musk-artificial-intelligence-open-ai-business-news-today-34802/#]/Kankee

The mission statement on the Open AI website says: “Discovering and enacting the path to safe artificial general intelligence.” So which is it: discovering (i.e., inventing something new) or protecting people from a category of human activity they call “artificial general intelligence”? Like Albert Einstein’s theory of relativity, it sounds as if there are two distinct things: general and special. Artificial general intelligence (AGI) is what **threatens humanity** because it can potentially be applied for any purpose. Instead of conveniently solving specific problems, it could have a direct **impact on the way humans understand** the world or rather think that they understand the world. Here is one [definition](https://www.extremetech.com/extreme/275768-artificial-general-intelligence-is-here-and-impala-is-its-name) of artificial general intelligence: “AGI is a single intelligence or algorithm that can learn multiple tasks and exhibits positive transfer when doing so, sometimes called meta-learning,” leading to “recursive self-improvement.” It’s the machine that improves, not the people who benefit from it. Only the machine will “know” what’s going on and how it got there. But the outcome will seem logical because it is logical, which of course doesn’t mean it’s true. It proves consistent with the patterns it finds and analyses the data it receives. With theoretically unlimited data, it will keep learning and applying a logic that may derive either from the initial algorithm, created by humans, or some extrapolation justified by the data. **Machine learning is the ultimate example of hyperreality**. Because it comes from a process rather than the behavior of real things, even if it takes into account their behavior, its results will end up looking as credible as reality — **simulating reality** without being reality. That explains why a hyperreal human operator like Elon Musk is eager to affirm his fear of the toolbox he himself is investing in and why we should understand that his fear justifies the investment. This positions him and his enterprise as a benefactor of humanity since its aim is to identify and presumably prevent evil uses of the tools. Open AI has cast itself in a role similar to that of the military instructor teaching recruits to “know the enemy.” Musk is our guide to the future, putting us in the role of the [hitchhikers of the galaxy](https://www.goodreads.com/book/show/386162.The_Hitchhiker_s_Guide_to_the_Galaxy).

#### AI systems destroy human life from media content, accelerating hyperreality- perpetuating a demon of simulacra that turns human lives into non-events

**Makris 22** [Spiros Makris, Associate Professor of Political Theory at the University of Macedonia and a Visiting Research Fellow at the Centre for Rights and Anti-Colonial Justice at the School of Global Studies at the University of Sussex, UK, 10-7-2022, “Digital Virulence and Post-Truth in Light of Baudrillard’s Science-Fiction Theory of Pataphysics,” E-International Relations, https://www.e-ir.info/2022/10/07/digital-virulence-and-post-truth-in-light-of-baudrillards-science-fiction-theory-of-pataphysics/#google\_vignette]/Kankee

Baudrillard reads the extreme phenomena of simulacra in late modernity drawing his inspiration from Marshall McLuhan’s axiom that medium dominates over the message (Baudrillard, 1996, 52). Message no longer expresses social ontology. Message is the medium itself (Baudrillard, 2019, 171–195). ‘From medium to medium’, he writes, the real is volatilized; it becomes an allegory of death, but it is reinforced by its very destruction; it becomes the real for the real, fetish of the lost object – no longer object of representation, but ecstasy of denegation and of its own ritual extermination: the hyperreal (Baudrillard 1983, 141–142). Undoubtedly, this is a turning point in his analysis to the extent that although the ideological usage of digitality varies, bringing to light a heterogeneous multitude of individual and collective attitudes and behaviours, many of them are absolutely violent and racist. For example: online vitriol, lack of civility in the cybersphere, and anti-feminist rhetoric in social media (Jane 2014; Cole 2015). For Baudrillard, medium has the pervasive force to homogenise everything. In fact, digital simulacrum transforms messages into pure operations, without any real connection (Baudrillard 1983, 100). By digitising social, political and cultural activity, IT, **AI**, mass media, screens and communication networks evacuate human life from content. Digital instantaneity, i.e. ‘real time’, decodes and denervates any referent. Digitality distorts and controls meaning (Baudrillard 1983, 115–123). Masses are absorbed into a hermetically closed global virtual circuit that reproduces them as a silent object sign. But, as mentioned above, this is half the story. While mass media inactivates masses – at the same time, the masses block mass media and global politics. It is worth pointing out here that the mediatised hyperreal world, composed of digital screens and simulacra, does not function as a mirror of truth and lies, but as a transparent and depthless void that devours every negativity and singularity (Baudrillard 1993, 111–174). Mass media is not a rational mechanism of verification and falsification. Instead, it works as an operational system that **overproduces** digital virulence in ‘real time’. Having transformed masses into an enormous brain, without body and blood, mass media circulates in it the virtual viruses of pure simulacra on a 24/7 basis (Baudrillard 2000). Despite the fact that Covid-19 brought to the fore the medical aspect of viruses, for Baudrillard, digital virulence stands in the very heart of late capitalism. It is no coincidence that he likens the ecstasy of communication to the situation of a cancerous metastasis. The excess of information **destroys** every possibility of meaning. The obscenity and promiscuity of mass media and digital screens nullifies every possibility to distinguish truth from falsehood. ‘The contradictory process of true and false’, he writes, ‘of real and the imaginary, is abolished in this hyperreal logic’ of mass media montage (Baudrillard 1983, 122). The exhaustion of reality murders the Real via the fatal strategy of perfect crime (Baudrillard 2008, 18). Nonetheless, it is important to remember that digitality means complicity. The masses participate in a post-orgy condition in which lies, fake news and trolling are not but the obese expression of a pornographic world (Happer et al. 2019). Digitality turns capitalism into a gigantic stage, where humanity exists performing in ‘real time’. As in a reality show, masses can only live through screen and simulacra. However, this is in vain. Digital hyperreality evacuates any message, distorts truth and false and corrupts axiological judgment. Only mass media, networks and platforms can deliver the pretext of objectivity. In the postmodern and posthuman world, true and false are **empty** signifiers that are floating in the depthless space of global digitality. As pixels, they unstoppably change in a dizzying void full of bizarre simulacra. Paraphrasing Heraclitus, it could be said that everything is given birth through a digital war (Merrin 2019). Following in Baudrillard’s footsteps, Merrin (2005) asserts that the digital screen ‘is killing the art of symbolic exchange’. In other words, the electronic **demon of simulacra** turns human life into a delirious spectacle of non-events. Reversibility principle, reality, and the political economy of TV fantasies

#### Hyperreality disintegrates truth, meaning, boundaries, and reality- marking humanity as calculable while plunging them into hyper-schizophrenia

**Pasco 16** [Marc Oliver D. Pasco, Assistant Professor of Philosophy in Ateneo de Manila University, 2016, “From Objects to Being and Beyond: Situating the Crisis of Reason Within the Bounds of a Hyperreal Interpretation of Contemporary Media Society.” International Journal of Baudrillard Studies, <https://baudrillardstudies.ubishops.ca/from-objects-to-being-and-beyond-situating-the-crisis-of-reason-within-the-bounds-of-a-hyperreal-interpretation-of-contemporary-media-society/>]/Kankee

Every characteristic thus elevated to the superlative power, caught in an intensifying spiral—more true than the true, more beautiful than the beautiful, more real than the real (hyper)—is assured a vertiginous effect that is independent of all content or specific quality, and which presently has tendency of being our only passion. The passion of intensification, of escalation, of mounting power, of ecstasy, of whatever quality so long as, having ceased to be relative to its opposite (the true to the false, the beautiful to the ugly, the real to the imaginary), it becomes superlative, positively sublime as if it had absorbed the energy of its opposite (Ibid.: 189-90). We exist in a world of excess. The production and reproduction of objects, information and ways of life happens in **hyper-speed** due largely to the availability of virtual platforms of transportation, communication and exchange. The production and reproduction of images, the supposed mirrors of the real has taken on a life of its own, pushing mankind to a state of schizophrenic hyper-awareness. The fulcrum of **reality** which was once embedded in the indubitable foundation of subjective certainty **disintegrates** and gives way to the insurmountable momentum of hyperreality. Today, the subjective and objective poles of epistemological knowledge implode into a pure screen of the obscene, the fractal, the logic of the code. The illusion projected on to the image rendering it a simulation of the real to the mind of the subject has transcended the limits of its intended definition. With the enigmatic coagulation of the real and the virtual in technological media society, copies no longer signify anything beyond themselves. **Reality and simulacra become indistinguishable**. The virtual’s revelation of the fractal translates Being itself into a code, a replicable series of signifiers that have lost their respective signifieds, rendering subjective belief and knowledge moot and to a certain extent, pretentious. As Baudrillard says, “It is no longer a question of imitation, nor of reduplication, nor even of parody. It is rather a question of substituting signs of the real for the real itself; that is, an operation to deter every real process by its operational double, a metastable, programmatic, perfect descriptive machine which provides all the signs of the real and short-circuits all its vicissitudes” (Ibid.: 170). Simulacra represents the obscene, it reveals the genetic reproducibility of the real, which**implodes** the traditional conceptualization of **what it means to be**. Through the translation of the metaphysical into the fractal, the holographic derivation of codes and models of reality through technological media dislodges the ground of objectivity from the realm of subjectivity and transfers all its power into the territory of pure simulacra. It’s not that there is nothing real. It’s that there too much and too many of it. As Horrocks explains, “The result of this absorption of illusion by technological reality is that credibility—a property of objects and images—has taken the place of belief—a property of ourselves as subjects. We judge events by their proximity to their code or model, rather than by some humanistic or metaphysical principle” (Horrocks, 2000: 35-6). The question concerning the meaning of being, pursued and seemingly resolved by modernity gave man the object. In the contemporary age of media technology, the question concerning the meaning of being, placed within the context of hyperreality, is imploded by the sheer iterability and density of the equiprimordiality of possible responses to the question, relegating the fundamentality of the question into perhaps a more thought-worthy sort of oblivion. V. Conclusion The age of mirrors, we might say, reached its historical **tipping point** in the age of modernity. The progress of scientific knowledge derives from the auto-assimilative movement of consciousness that has finally become aware of itself as the ground for objectivity. The Cartesian self-certain cogito, by way of the logic of metaphysical transitivity, reproduces itself in its own ideas, projected as the very foundation of existence itself. The mathematization of reality sets up the parameters of knowledge in advance of unconcealment. The realm of truth, primordially interpreted by the ancients as aletheia, is engineered in accordance with the subjective directives of measurability and calculability. Heidegger reckons that this epochal transmutation of the truth of Being from unconcealment into objectivity is an oblivion of Being itself; but it is a forgetfulness that is destined by the very movement of Being in history. In other words, the eclipse of the realm of unconcealment by the object of modern science is heralded and is made possible by the essence of truth itself—the realm of unconcealment must retreat from the horizon of what is in order for appearing to happen. This is what principally constitutes the age of objects. The object absorbs the idolatrous gaze of the subiectum and functions as the primordial mirror of the self. The self-certain ego of modernity propagates itself by perpetually guaranteeing its personal ontological and epistemological certainty in its objects. And since there is no other logical way about it, in this regard, man’s relationship to Being is reduced to measurability and calculability, and truth becomes an attribute of representations. In the age of contemporary technological media however, objects begin to vanish into pure visibility. The mirror, as it were, is rendered opaque by its obscene exposure to itself as mirror as it gravitates towards the hyperbolic momentum of hyperreality. The exponential multiplication of mirrors cancels out the polarizing effect of subjective alienation, which used to serve as the ground for objectivity, and delivers the subject into a state of schizophrenic hyper-awareness. Once the fractal, genetic code of reality became installed into the logic of the social by media, the reproducibility of the real terminates in its excess. As opposed to its modern version, the contemporary object’s absorption of the light of reason depletes the subjective foundations of certitude and reveals itself as pure simulacrum. The incessant exposure of all aspects of the real in our age overrides the principles of measurability and calculability and proceeds to hypertrophically reveal all our measures and calculations to ourselves as the logical offshoot of the very reproducibility of objects, images, ideas and events. The digital code of the real, in other words, takes on a life of its own and perpetuates its being through the momentum of technological media. In this respect, the objectivity of the subjective perspective is nullified and the obscene exposure of the real metastasizes into an operational function of hyperreality. As Douglas Kellner explains, “The narcoticized and mesmerized media-saturated consciousness is in such a state of fascination with image and spectacle that the concept of **meaning itself** (which depends on stable boundaries, fixed structures, shared consensus) **dissolves**” (Kellner, 1994: 9). The real is no longer the rational, as it were. Being itself, reduced to the logic of the fractal and the digital implodes the boundaries of metaphysical and epistemological certitude and transforms reality from being a function of consciousness into that which transcends consciousness by way of obscenity. The hyper-aware schizophrenic, as Baudrillard’s depiction of contemporary man suggests, is not seen as the locus of meaning, but a mere conduit for the functional operation of digital hyperreality. As Foster expounds, “No more hysteria, no more projective paranoia, properly speaking, but this state of terror proper to the schizophrenic: too great a proximity of everything, the unclean promiscuity of everything which touches, invests and penetrates without resistance, with no halo of private protection, not even his own body, to protect him anymore” (Foster, 1998: 132). We now exist in a world where life is certainly stranger than fiction. The clear line which used to divide the real from the false, the beautiful from the ugly and good from evil has been blurred, if not **totally expunged** by the obscene—the more real than the real, the more beautiful than the beautiful. As we are constantly becoming more and more exposed to every aspect of Being in the hope of recovering from the alienation bequeathed to us by the epistemological resolve of modernity, the more Being is eclipsed by the hyperreal. Media, in all its forms, delivers awareness and information. The question, however, is whether we still carry autonomy with respect to what we can or cannot know or to what we want or do not want to see. The Heideggerian narrative of oblivion placed man in the threshold of a hopeful disposition of openness and thought. Baudrillard’s story, however, seems not as promising. The condition for the possibility of genuine thought with regard to one’s standing in relation to Being, in my estimation, requires at the very least some degree of self-awareness borne from the perspective of historical and hermeneutic distanciation. Heidegger, speaking from the point of view of someone witnessing the birth of the nuclear age, named this disposition, *releasement.* In the age of hyperreality, however, historical and hermeneutic distanciation seems to be next to impossible. As the simulacrum of the real perpetuates itself in the instantaneously reproducible obscene, everything becomes too real and too immediate for us to rationally gather ourselves and process. As Baudrillard famously opines, “The day that there is a real war you will not even be able to tell the difference” (Baudrillard, 2001: 170). This, I think, is the contemporary challenge for genuine thought. In an age where we can no longer properly distinguish the real from the false, how must we think? In an age where history is recycled and events disappear into internet links, how must we narrate our stories? In an age where nothing is no longer secret, how must we seek? These questions, I think, are very difficult to answer. It is at this point, I think where Heidegger’s famous line from his 1966 *Der Spiegel* interview rings most true. Perhaps only a God can save us now.

## Negative

### Contention 1: Innovation

#### Permissionless innovation is strong now - Congress is too slow

Thierer 21 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 1-6-2021, "The Future of Innovation: Is This the End of Permissionless Innovation?", Discourse Magazine, https://www.discoursemagazine.com/p/the-future-of-innovation-is-this-the-end-of-permissionless-innovation]/Kankee

Many independent telephone companies existed across America before AT&T’s leaders cut sweetheart deals with policymakers that tilted the playing field in its favor and undermined competition. With rivals hobbled by entry restrictions and other rules, Ma Bell went on to enjoy more than a half century of stable market share and guaranteed rates of return. Consumers, by contrast, were expected to be content with plain-vanilla telephone services that barely changed. Some of us are old enough to remember when the biggest “innovation” in telephony involved the move from rotary-dial phones to the push-button Princess phone, which, we were thrilled to discover, came in multiple colors and had a longer cord. In a similar way, the impending close of the permissionless innovation era signals the twilight of technological creative destruction and its replacement by a new regime of political favor-seeking and logrolling, which could lead to innovation stagnation. The CEOs of the remaining large tech companies will be expected to make regular visits to the halls of Congress and regulatory agencies (and to all those fundraising parties, too) to get their marching orders, just as large telecom and broadcaster players did in the past. We will revert to the old historical trajectory, which saw communications and media companies securing marketplace advantages more through political machinations than marketplace merit. Will Politics Really Catch Up? While permissionless innovation may be falling out of favor with elites, America’s entrepreneurial spirit will be hard to snuff out, even when layers of red tape make it riskier to be creative. If for no other reason, permissionless innovation still has a fighting chance so long as Congress struggles to enact comprehensive technology measures. General legislative dysfunction and profound technological ignorance are two reasons that Congress has largely become a non-actor on tech policy in recent years. But the primary limitation on legislative meddling is the so-called pacing problem, which refers to the way technological innovation often outpaces the ability of laws and regulations to keep up. “I have said more than once that innovation moves at the speed of imagination and that government has traditionally moved at, well, the speed of government,” observed former Federal Aviation Administration head Michael Huerta in a 2016 speech. The same factors that drove the rise of the internet revolution—digitization, miniaturization, ubiquitous mobile connectivity and constantly increasing processing power—are spreading to many other sectors and challenging precautionary policies in the process. For example, just as “Moore’s Law” relentlessly powers the pace of change in ICT sectors, the “Carlson curve” now fuels genetic innovation. The curve refers to the fact that, over the past two decades, the cost of sequencing a human genome has plummeted from over $100 million to under $1,000, a rate nearly three times faster than Moore’s Law. Speed isn’t the only factor driving the pacing problem. Policymakers also struggle with metaphysical considerations about how to define the things they seek to regulate. It used to be easy to agree what a phone, television or medical tracking device was for regulatory purposes. But what do those terms really mean in the age of the smartphone, which incorporates all of them and much more? “‘Tech’ is a very diverse, widely-spread industry that touches on all sorts of different issues,” notes tech analyst Benedict Evans. “These issues generally need detailed analysis to understand, and they tend to change in months, not decades.” This makes regulating the industry significantly more challenging than it was in the past. It doesn’t mean the end of regulation—especially for sectors already encumbered by many layers of preexisting rules. But these new realities lead to a more interesting game of regulatory whack-a-mole: pushing down technological innovation in one way often means it simply pops up somewhere else. The continued rapid growth of what some call “the new technologies of freedom”—artificial intelligence, blockchain, the Internet of Things, etc.—should give us some reasons for optimism. It’s hard to put these genies back in their bottles now that they’re out. This is even more true thanks to the growth of innovation arbitrage—both globally and domestically. Creators and capital now move fluidly across borders in pursuit of more hospitable innovation and investment climates. Recently, some high-profile tech CEOs like Elon Musk and Joe Lonsdale have relocated from California to Texas, citing tax and regulatory burdens as key factors in their decisions. Oracle, America’s second-largest software company, also just announced it is moving its corporate headquarters from Silicon Valley to Austin, just over a week after Hewlett Packard Enterprise said it too is moving its headquarters from California to Texas—in this case, Houston. “Voting with your feet” might actually still mean something, especially when it is major tech companies and venture capitalists abandoning high-tax, over-regulated jurisdictions. Advocacy Remains Essential But we shouldn’t imagine that technological change is inevitable or fall into the trap of thinking of it as a sort of liberation theology that will magically free us from repressive government controls. Policy advocacy still matters. Innovation defenders will need to continue to push back against the most burdensome precautionary policies, while also promoting reforms that protect entrepreneurial endeavors. The courts offer us great hope. Groups like the Institute for Justice, the Goldwater Institute, the Pacific Legal Foundation and others continue to litigate successfully in defense of the freedom to innovate. While the best we can hope for in the legislative arena may be perpetual stalemate, these and other public interest law firms are netting major victories in courtrooms across America. Sometimes court victories force positive legislative changes, too. For example, in 2015, the Supreme Court handed down North Carolina State Board of Dental Examiners v. Federal Trade Commission, which held that local government cannot claim broad immunity from federal antitrust laws when it delegates power to nongovernmental bodies, such as licensing boards. This decision made much-needed occupational licensing reform an agenda item across America. Many states introduced or adopted bipartisan legislation aimed at reforming or sunsetting occupational licensing rules that undermine entrepreneurship. Even more exciting are proposals that would protect citizens’ “right to earn a living.” This right would allow individuals to bring suit if they believe a regulatory scheme or decision has unnecessarily infringed upon their ability to earn a living within a legally permissible line of work. Meanwhile, there have been ongoing state efforts to advance “right to try” legislation that would expand medical treatment options for Americans tired of overly paternalistic health regulations. Perhaps, then, it is too early to close the book on the permissionless innovation era. While dark political clouds loom over America’s technological landscape, there are still reasons to believe the entrepreneurial spirit can prevail.

#### The aff causes uncertainty and delays that destroys broader innovation culture

Thierer 23 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 03-2023, “Getting Al Innovation Culture Right,” R Street Policy Study, https://www.rstreet.org/wp-content/uploads/2023/03/Final-Study-No.-281.pdf]/Kankee

\*text converted via OCR

Getting Al Innovation Culture Right Technological revolutions are born from innovation cultures that reward risktaking; investment; and the free movement of people, products, capital and ideas.\*® Every restriction that limits those things—or that even creates confusion and uncertainty about them—will undermine a nation’s innovation culture and its prospects for prosperity. If the United States is to lead the computational revolution and repeat its success in the digital revolution, it is essential that policymakers revisit and reembrace the original principles and policies. Importantly, policy for emerging technologies must not be fear-based or built on a premise of worst-case thinking about the future. Technology critics tend to look at new Al innovations and see only potential problems, all of which they want to preemptively address before new algorithmic technologies are encouraged. This can be thought of as “regulation by hypothesis,” or policymaking by worst-case hypotheticals without any regard to the opportunity cost of proposed regulatory restrictions.\*” Concerns about emerging technologies like Al should be taken seriously, of course, but it is equally important that humans be able to flourish and enjoy the fruits of innovation. There is a compelling interest in ensuring that innovations are developed and made widely available to society. Technological critics often casually assume that important innovations will just magically come about, and they jump ahead to ponder all the ways they believe we will need to control the future. However, there is no need to worry about the future if inventors cannot even create it first. The danger exists that policy for algorithmic systems could be formulated in such a way that innovations are treated as guilty until proven innocent—i.e., a precautionary principle approach to policy—resulting in many important Al applications never getting off the drawing board. If regulatory impediments block or slow the creation of life-enriching, and even life-saving, Al innovations, that would leave society less well-off and give rise to different types of societal risks.®® Another technology policy priority is to avoid replicating the regulatory mistakes of the past. We know what does not work because we have a full century of well-documented failures to learn from, and we know what does work thanks to the recent successes the United States has enjoyed using a very different policy approach for the internet and ecommerce. Put simply, the United States must reject the precautionary principle-based thinking and policies of the old analog era and instead reembrace the permissionless innovation vision of the digital era. The Two Great Lessons of Analog-Era Tech Policy The history of the United States’ analog-era policy regime offers two crucial lessons for technological governance more generally: 1. First, no matter how well intentioned any rules may be, preemptive prior restraints on innovative activities will generate many different costs and unintended consequences. Just because someone claims that something is “in the public interest” does not automatically mean it is. Real-world results matter more than good intentions. 2. Second, we cannot pursue better market-driven ways to address important policy goals when heavy-handed regulation makes them difficult or impossible. The touchstones of good policy are humility and flexibility. Consider how both lessons were evident in the past—but also largely ignored—in the context of wireless spectrum and universal service policy. ® Spectrum policy: In the 1950s and 60s, an economist who proposed property rights and auctions to better allocate wireless spectrum was laughed out of the room at a 1959 FCC hearing.® At the time, it was thought to be in the public interest to assign spectrum through a top-down licensing regime that tightly limited the use and sale of any wireless service. In essence, it was an inflexible “zoning” regime for spectrum use. While it is impossible to know exactly how much earlier robust, nationwide wireless markets might have developed had policymakers heeded economists’ advice to tap the power of market incentives, it is likely that the opportunity costs of this policy miscalculation were significant. Investments in alternative communications and media platforms, services and devices were delayed for decades until the FCC liberalized spectrum markets and used auctions to allocate wireless services in the 1990s. To appreciate the true costs of this decision, imagine if the FCC would have possessed authority over the computer sector during the 1950s and used its authority to dictate that only vacuum tube mainframes were “in the public interest” and would, therefore, be federally licensed and regulated. Transistorized computers and the personal computer revolution would have likely been delayed significantly had such a regulatory regime been in place because massive mainframes were thought to be the only machines capable of serious computational tasks. Universal telephone service: Another example of good intentions gone wrong involves universal service. Ensuring that the public was connected to basic telephone service was a worthy policy goal over the last century, but it did not need to be limited by inflexible, highly inefficient, top-down regulatory mandates and controls. Instead, policymakers could have opted to “voucherize” universal service assistance, allowing consumers to shop around for telecom and media service alternatives using a means-tested government voucher.®® Unfortunately, instead of passing out pro-competitive vouchers to generate pro-competitive incentives, governments passed out local monopolies and then demanded that those firms always offer the community basic service. Imagine if, in the name of ensuring that every community had low-cost food, the first grocery store or restaurant in town had to serve everyone the same (price-controlled) food in exchange for protection from any potential competitors that followed. That would be a highly inefficient way to pursue such goals, yet it was the law of the land for almost a full century for telecommunications in the United States. Things could have worked differently with vouchers. Just as policymakers long ago adopted food stamps to give people the flexibility to buy the food they wanted from the store they wanted, policymakers could have similarly used means-tested “phone stamps” to let households shop around for their communications or media needs. Unfortunately, even now, America is still struggling to find efficient ways to provide broadband access to underserved individuals when pro-competitive solutions could be implemented.®\* The common themes in both of these examples were mandates over markets; top-down regulatory decision-making over bottom-up, consumer-driven processes; and policy rigidity over flexible experimentation. These policy choices restricted entrepreneurialism, competition and consumer choice in myriad ways. In short, they created a suboptimal innovation culture that had to be abandoned to unlock the full potential of the U.S. ICT sector. The United States turned an important corner when policymakers moved away from that regime to close the 20th century and embraced a fresh approach for computing, data services and the digital economy. The defining feature of the new approach was an embrace of permissionless innovation, and a corresponding rejection of the precautionary principle as the default for ICT policy. Generally speaking, flexible, bottom-up, consumer-driven governance beats technocratic, top-down regulation. The United States did not need a grandiose regulatory plan or over-arching bureaucracy to guide the development and growth of the internet. In fact, digital entrepreneurialism and online innovation flourished precisely because the U.S. did not adopt such mandates or technocratic agencies.” Had the United States created a Federal Computer Commission or a National Internet Agency, the resulting red tape burdens would have left us no better off than Europe, where mountains of paperwork compliance requirements resulted in a staggering loss of competitive advantage.® It is difficult to name any leading global information technology companies based in Europe because heavy-handed regulations and overlapping bureaucracies kneecapped digital entrepreneurs and forced many European innovators and investors to jump the Atlantic and launch their ideas here instead.\* Critics will claim that many unforeseen privacy and security problems developed due to the rise of the internet and digital networks. That is true, and we are still devising solutions to many of those issues. But we should not fool ourselves into believing we could have solved all of these problems preemptively through regulatory mandates—at least not without fundamentally stunting the development of digital technologies the same way telecom and media innovation and competition were stifled in the previous century by overbearing regulatory mandates. We should work through challenges as they come at us, but the right policy default for the internet and for Al continues to be “innovation allowed.” Entrepreneurs and their creations must be treated as innocent until proven guilty. Table 1 illustrates how this approach to innovation policy contrasts with the precautionary principle oriented vision of technological governance. When problems arise, there exist many ex-post (responsive) flexible governance remedies including various common law solutions (torts, class actions, contract law, etc). And there are many regulations already, including the recall authority possessed by many regulatory agencies as well as various federal and state consumer protection policies.” To reiterate what the Framework for Global Electronic Commerce properly recommended, “where governmental involvement is needed, its aim should be to support and enforce a predictable, minimalist, consistent and simple legal environment for commerce.” Biden’s Al “Bill of Rights”—The Beginning of Burdensome Regulations? The policy choices we make now will help determine which innovation culture the United States creates for Al and algorithmic innovations. The question is whether policymakers will reembrace the permissionless innovation vision that powered digital technologies and online services or revert to the permission-slip-oriented regulatory model of the analog era. I Some scholars note that evidence suggests, “we are moving away from permissionless innovation and toward the precautionary principle.”\*® Indeed, the Biden administration and many in Congress appear ready to reverse course and abandon the highly successful policy legacy of the permissionless innovation era.”” While the Obama and Trump administrations generally embraced the Clinton administration’s market-driven vision for ICT, many policymakers are currently floating more aggressive regulatory approaches for both existing digital platforms as well as new algorithmic systems.® With China and other countries In October 2022, the White House released a “Blueprint for an Al Bill of Rights” (A/ attempting to catch up to the United States on the algorithmic technology front, the United States must and Protect the Rights of the American Public” (Key Actions).\* This is part of a implement smart policy that supports innovation and avoids a burdensome new regulatory regime. Blueprint) and an accompanying list of “Key Actions to Advance Tech Accountability growing effort by the Biden administration to craft a broad-based governance framework for algorithmic systems. While some of the recommendations in the Al Blueprint and corresponding Key Actions document are quite amorphous and aspirational, several could be burdensome in practice and represent a major setback for U.S. efforts to be a leader in global Al competition. With China and other countries attempting to catch up to the United States on the algorithmic technology front, the United States must implement smart policy that supports innovation and avoids a burdensome new regulatory regime.” Importantly, however, the 73-page Al Blueprint begins with a disclaimer that the document “is non-binding and does not constitute U.S. government policy” and also “does not require compliance with the principles described herein.”’\* Thus, what follows in the document is merely a set of aspirational principles or suggested best practices, which is welcome news. If these best practices were to remain in the realm of “soft law”—i.e., unbinding and informal norms and standards—they might not be as constraining or burdensome in practice. The Al Blueprint dodges the question of whether these principles might be converted into formal policies, but an earlier Biden administration sketch of “a Bill of Rights for an Al-Powered World” suggested the need for “new laws and regulations to fill gaps,” and that “states might choose to adopt similar practices.””? With that framing in mind, the Al Blueprint opens by claiming that algorithmic systems are “unsafe, ineffective, or biased”; “deeply harmful”; “threaten the rights of the American public”; and “are used to limit our opportunities and prevent our access to critical resources or services.”” The Al Blueprint continues in this vein, repeatedly stressing possible dangers over potential opportunities.” Fear-driven policymaking It is certainly true that Al systems have their share of faults and potential dangers, just unflermlnes innovation culture. as every new technology poses certain risks. Some of these risks are real; others are hypothetical. Nevertheless, the Al Blueprint mostly stresses worst-case possibilities. This is an issue because fear-driven policymaking undermines innovation culture. Living in constant fear of worst-case scenarios—and premising public policy on them—means that best-case scenarios will never come about.” This, in turn, denies the public many potential benefits of technologies that may be delayed or kept off the market. Instead, the United States needs a more flexible governance vision for Al that rejects fear-based policymaking as its starting point. And we can look to the Clinton-era internet principles for a more positive innovation culture vision of algorithmic governance. Both the Biden administration’s Al Blueprint and the Clinton administration’s Framework contain five core principles, but those principles diverge considerably.” The Biden Al Blueprint focuses on affirmative obligations and constraints for Al innovators, whereas the Clinton Framework focuses on entrepreneurial freedoms.” P Consider the Al Blueprint’s first principle: “You should be protected from unsafe or ineffective systems.””® Although that initial premise is reasonable, the Al Blueprint goes on to advise, “[a]utomated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems.” Going further, the Al Blueprint suggests several other obligations for Al developers regarding this principle: \* “The public should be consulted in the design, implementation, deployment, acquisition, and maintenance phases of automated system development, with emphasis on early-stage consultation before a system is introduced or a large change implemented.” \* “Systems should undergo extensive testing before deployment.” \* “Before deployment, and in a proactive and ongoing manner, potential risks of the automated system should be identified and mitigated.” \* “In some cases, it may be appropriate for an independent ethics review to be conducted before deployment.” Although these best practices are not inherently objectionable, if they were to be translated into regulatory requirements improperly, they could give rise to a more convoluted process for algorithmic design based on highly subjective and potentially unforeseeable potential future risks. Other principles found in the report seek to limit data collection and create broad, open-ended requirements for algorithmic innovators to create opt-out or complaint procedures. Another of the five major principles relates to avoiding algorithmic discrimination, and the document suggests that this “should include proactive equity assessments as part of the system design” as well as a formal algorithmic impact assessment. Algorithmic audits and impact assessments have become increasingly popular in the field of Al governance, with many academics and others suggesting that they could be a useful tool for evaluating algorithmic design and functionality in a search for various risks.” Congressional lawmakers are floating many new laws to regulate Al in a top-down fashion.® The Algorithmic Accountability Act of 2022 would require that any large company that “deploys any augmented critical decision process” must undertake algorithmic impact assessments “to eliminate or mitigate, in a timely manner, any impact made by an augmented critical decision process that demonstrates a likely material negative impact that has legal or similarly significant effects on a consumer’s life.”®! The law mandates that firms file those audits with the Federal Trade Commission and creates a new Bureau of Technology within the agency to oversee the process. Legislators have proposed many other laws to regulate algorithms in an attempt to address broader social issues, such as hate speech, conspiracy theories and child safety, including the “Protecting Americans from Dangerous Algorithms Act.” Many academics have already promoted formal regulatory ideas and agencies like an “FDA for Algorithms,” a National Algorithmic Technology Safety Administration, or an Al Control Council, among other ideas.® Other academics have suggested modeling algorithmic audits and permitting procedures after the National Environmental Policy Act (NEPA), a 1970 law that requires formal environmental impact statements for major federal actions “significantly affecting the quality of the human environment.”®\* Many states have adopted similar regulatory requirements. NEPA is a troubling model for algorithmic audits because it is now widely acknowledged that the law slows progress on important societal goals and projects. Analysts have found that NEPA assessments, which were initially just a few pages long, today average more than 600 pages and include appendices that average over 1,000 pages.®\* These assessments now also take an average of 4.5 years to complete, and some take much longer.® This process has held up or even derailed many important infrastructure projects and clean energy initiatives. One NEPA expert notes that the law has also become highly politicized and “seems easily captured by small groups with strongly held opinions” who stand ready to block almost all progress on important projects, adding, “[b]ecause it adds cost and uncertainty to any new major project, NEPA is effectively a bias towards the status quo.”\*® The Atlantic notes that, “many people within the environmentalist movement are undermining the nation’s emissions goals in the name of localism and community input.”®’ If NEPA becomes a model for mandated algorithmic audits, Al innovation would similarly grind to a halt in the face of lengthy delays, formidable paperwork burdens and considerable compliance costs. Opponents of different forms of Al innovation would use the mandatory auditing process to slow or block important algorithmic advances, all in the name of “democratic input” that would not likely be all that democratic in reality. Instead, a small number of vociferous regulatory advocates and special interests would simply use the process to constantly veto new ideas and products. To reiterate, as abstract best practices, many of the A/ Blueprint’s recommendations are unobjectionable, including the notion that Al developers should regularly evaluate their algorithms for negative impacts. Using occasional voluntary audits might be one way of carrying out this objective. However, if the Al Blueprint’s principles come to inspire the passage of new laws like the Algorithmic Accountability Act, or if they encourage federal regulatory agencies to aggressively regulate under amorphous existing authority, or even if they encourage states to impose a patchwork of algorithmic mandates, the resulting regulatory system would open Al innovators to massive ex-ante compliance costs or ex-post liability threats. If such principles were imposed through a top-down, technocratic system, the United States would be walking down the path that the European Union (EU) has already charted with its regulatory regime for the data-driven economy, which decimated its information technology sector.?® The EU is now preparing to expand that regime through a new Artificial Intelligence Act, which creates a new European Artificial Intelligence Board that will enforce a complex system of algorithmic “conformity assessments” and impose steep fines for violations. Compliance costs will likely devastate small and medium sized enterprises and limit Al innovation across the continent.® The European Commission itself estimates that the mandate to set up the quality management systems required by the law will cost roughly $193,000-$330,000 upfront plus $71,400 in yearly maintenance cost.®® Toward Agile Governance for Al

#### The aff spillsover to stifle innovation in other industries out of fear of regulation

Thierer 23 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 03-2023, “Getting Al Innovation Culture Right,” R Street Policy Study, https://www.rstreet.org/wp-content/uploads/2023/03/Final-Study-No.-281.pdf]/Kankee

\*text converted via OCR

Conclusion: Embracing a Dynamic, Open Future As policymakers consider governance solutions for Al and computational systems, they should appreciate how a policy paradigm that stacks the deck against innovation by default will get significantly less innovation as a result. Innovation culture is a function of incentives, and policy incentives can influence technological progress both directly and indirectly. Over the last half century, “regulation has clobbered the learning curve” for many important technologies in the United States in a direct way, especially those in the nuclear, nanotech and advanced aviation sectors.'® Society has missed out on many important innovations because of endless foot-dragging or outright opposition to change from special interests, anti-innovation activists and over-zealous bureaucrats. Furthermore, when government leaders and other critics demonize Al and computational science, it discourages individuals from studying or pursuing careers in these fields. After all, few would want to try to operate in an innovation cage, constantly struggling to gain the freedom to experiment, when other sectors (or even nations) offer a more hospital environment. The Biden administration’s Al Blueprint could send a similar message. It reads more like a blueprint for aspiring tech critics and trial lawyers who hope to bottle up algorithmic innovations rather than helping to advance them. That is not the best way for a country to craft a positive innovation culture. Our nation’s policy toward Al, robotics and algorithmic innovation should embrace a dynamic future and the enormous possibilities that await us.

#### History proves positive innovation cultures drive tech development

Thierer 23 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 03-2023, “Getting Al Innovation Culture Right,” R Street Policy Study, https://www.rstreet.org/wp-content/uploads/2023/03/Final-Study-No.-281.pdf]/Kankee

\*text converted via OCR

Executive Summary Public and political interest is intensifying in artificial intelligence (Al), machine learning and robotics. As these technological capabilities advance, legislative and regulatory proposals for algorithmic systems will grow alongside them. Public policy will play a crucial role in shaping the so-called “computational revolution.” To ensure that the United States can be a global leader in advanced technology sectors, we must create a policy innovation culture that encourages and rewards the entrepreneurial spirit of the American people. The danger exists that the United States could adopt the opposite approach, locking entrepreneurs and investors in an “innovation cage” that constrains their growth opportunities. Ultimately, policymakers must make a choice between two general policy defaults that will govern most algorithmic systems: the precautionary principle or permissionless innovation. Under the highly risk-averse precautionary principle approach, algorithmic innovations would essentially be treated as guilty until proven innocent, a legal standard generally shunned as unfair to individuals. Under the permissionless innovation approach, Al entrepreneurism is generally given a green light and treated as innocent until proven guilty, ensuring that people are mostly at liberty to create new things. This paper explores the dangers of adopting a highly regulatory approach and recommends continuing with the more permissionless approach to policy that helped spawn the digital revolution and made U.S. tech companies global powerhouses. Although some safeguards will be needed to minimize certain Al risks, a more flexible, bottom-up (i.e., less regulated) governance approach can address these concerns without creating overbearing, top-down (i.e., more regulated) mandates, which would hinder algorithmic innovations. The ramifications of this policy choice are significant because Al and algorithmic systems play an important role in the United States’ global competitive advantage and relative geopolitical power. With China becoming a major competitor in advanced information technology sectors and other nations racing to be at the The United States must create a positive forefront of the unfolding computational revolution, the United States must create a positive innovation culture if it hopes to prosper economically and ensure a safer, secure technological base. The Importance of Innovation Culture Technological innovation has been a key driver of improvements in human wellbeing throughout history, producing greater economic growth, worker opportunities and societal choices.1 Additionally, nations that have created a more positive innovation culture have enjoyed greater technological advancement than those that have not.2 Key elements of this type of culture include “attitudes towards innovation, technology, exchange of knowledge, entrepreneurial activities, business, uncertainty,” and related activities.3 The foundation of a positive innovation culture is a dynamic, open economy that encourages new entry; entrepreneurialism; continuous investment; and the free movement of goods, ideas and talent.4 Public policy has a strong influence on these prerequisites.5 The most basic lesson of economic history can be simply stated: You will only get as much innovation as you allow.6 Whenever new technologies challenge the existing economic or societal status quo, opponents to change emerge.7 All too often, these anti-technological forces use public policy to erect barriers to innovation and entrepreneurialism, locking in archaic rules or systems that benefit incumbent companies and other special interests.8 This is the “innovation cage” problem, in which these forces lock down entrepreneurial activities by default. A positive innovation culture requires that individuals have the freedom to try new things without needing bureaucratic permission at every juncture. The rise of the internet and the explosive growth of the digital economy in the past 20 years demonstrates this. Twenty-five years ago, the Clinton administration created a bold vision for internet governance that allowed the United States’ information and communications technology (ICT) sectors to break out of the innovation cage that had constrained the entrepreneurial spirit and economic potential of the nation.9 Once unleashed, the U.S. ICT sector became “a growth powerhouse” that drove “remarkable gains, powering real economic growth and employment.”\*® That policy vision continues to resonate today as the United States considers how to promote and govern Al, machine learning, robotics and algorithmic systems more generally. The same principles that powered the digital revolution can now drive the computational revolution—but only if the nation gets its innovation culture right. Unfortunately, many policymakers already appear to be heading in the wrong direction regarding important emerging technologies. For example, last year, the Biden administration released an “Al Bill of Rights” that—while not yet a formal regulatory agenda—represents a fear-based model of technology policymaking.\*\* The effort foreshadows what could become a precautionary principle-based policy regime for Al and the computational technologies of the future. If such an approach were to become the basis of Al policy, the resulting legal standard for many algorithmic systems would become what some pro-regulatory law professors refer to as “unlawfulness by default.”\*? In other words, Al entrepreneurs and their innovations would essentially be treated as guilty until proven innocent. Concerningly, unlawfulness by default could also result in technological stagnation by default. With this issue in mind, we outline in this paper how highly precautionary policy approaches have historically created innovation cages for algorithmic entrepreneurs and investors. We discuss extensively the value of rejecting such approaches and reembracing and extending the Clinton administration’s sensible innovation culture for digital technologies, which is rooted in the idea of permissionless innovation—the notion that entrepreneurs and innovations are innocent until proven guilty.!\* Importantly, we also offer practical strategies that policymakers can consider to help the United States create the positive innovation culture needed to unlock more economic opportunities and meet the growing global competition from China and other nations in cutting-edge emerging technology sectors. A Brief History of the United States’ Disastrous Analog-Era Policy Regime

#### The permissionless innovation precedent is key to global economic growth and tech – the aff causes uncertainty, innovation fears, and follow-on regulation

Moutii 23 [Mohamed Moutii, research associate with the Arab Center for Research and the Montreal Economic Institute, 10-14-2023, "The Future of Innovation in the United States: Permissionless or Regulated?", Econlib, https://www.econlib.org/the-future-of-innovation-in-the-united-states-permissionless-or-regulated/]/Kankee

The last decades in the United States have been marked by what has been called the era of “permissionless innovation.” From the early days of the internet to the rise of mobile computing, American entrepreneurs and innovators have been free to experiment and create new products and services without being hampered by excessive regulations or government interference. In a free market, individuals and businesses can freely innovate, compete, and succeed based on their merits without government favoritism or intervention. The culture of permissionless innovation in the US exemplifies this free-market ideal, enabling entrepreneurs and innovators to pursue their ideas and compete in the market freely. This strategy has fostered significant growth and innovation in the tech industry and other sectors. The achievements of American-founded companies like Amazon, Google, and Facebook exemplify the power of the permissionless innovation culture in fostering economic growth and development. Yet recent antitrust suits may indicate an end to that era. Many factors push the US towards greater public involvement, oversight, and stricter regulations. A Clash of Approaches: Permissionless Innovation vs. the Precautionary Principle Permissionless innovation is a term that refers to the idea that anyone should be able to innovate without having to seek permission from a government or other authority. It means that, as a general rule, experimentation with new technologies and business models should be allowed unless there is a strong argument that it will cause significant harm to individuals. Conversely, the precautionary principle suggests that innovations should be limited or prohibited until their inventors can verify their lack of harm toward individuals, groups, specific entities, cultures, or existing laws, norms, or traditions. This principle places the responsibility of proof on those proposing or engaging in an activity to demonstrate its safety rather than on those expressing concerns about potential risks. The United States is often associated with a culture of permissionless innovation. The US has traditionally embraced a more of laissez-faire approach to regulation, prioritizing economic growth, competitiveness, and technological advancement. This approach can be seen in various aspects of the US system, such as the relatively minimal regulatory burden on startups and technology companies, the freedom given to entrepreneurs to experiment and invent, and the emphasis on market forces to drive innovation. As a result, the US has been known for fostering groundbreaking innovations in industries like technology, biotech, and finance. In the mid-1990s, the Clinton administration made a wise choice. They declared the internet a “market-driven area,” not regulated, with limited government involvement only to support and enforce a predictable, minimalist, consistent, and simple legal environment. This policy allowed a new generation of creative minds to explore this frontier for business and commerce. This approach led to the internet’s success, resulting in a surge of innovation. Today, the US is home to the most innovative tech firms, hosting vibrant internet-based companies and bringing countless benefits to consumers and small businesses. On the other hand, Europe leads in adopting the precautionary principle in its regulations and decision-making. European regulatory frameworks prioritize consumer protection, public health, and environmental preservation. This approach can lead to stricter regulations, mandatory impact assessments, and more rigorous safety standards. It has been applied in various areas, including genetically modified organisms (GMOs), chemicals, pharmaceuticals, and emerging technologies. Nonetheless, inadequate regulation can result in unanticipated outcomes, including privacy worries, market monopolies, and detrimental societal effects. Overregulation hampers advancement, deters investments, and undermines the competitiveness of European industries. Europe’s precautionary principle culture has led to increased costs, market distortions, and a risk-averse approach to innovation in Europe. This principle, employed in Europe when faced with scientific uncertainty, has had numerous unintended consequences. For instance, the EU’s ban on genetically modified food crops has spiked food prices while reducing agricultural productivity, making it challenging for European farmers to compete globally. Additionally, it has led to a decrease in innovation. Companies are less likely to invest in new products and technologies if they are unsure whether they can market them in the EU. This situation has adversely impacted the EU economy. Precautionary measures enforced through stringent regulations and requirements can increase expenses for enterprises, ultimately affecting consumers. Moreover, they can discourage investments and hinder progress in specific industries. Every year, Boston Consulting Group (BCG) has compiled a list of the world’s most innovative companies. Notably, American tech companies have consistently dominated this list. In 2013, seven of the top ten most innovative companies were based in the United States. Despite increasing global competition, the 2023 list still features 16 U.S.-based giants among the top 25, including Amazon, Alphabet, Intel, Microsoft, Apple, Facebook, Space X, and Tesla. In stark contrast, European digital tech companies were conspicuously absent from this prestigious list back in 2013 and have a rare presence in the 2023 list, as mentioned in the chart below. Only the US and China have expanded their portion since 2013, displacing companies from European nations such as Germany, the UK, and Italy. Notably, there have been noticeable decreases in the presence of Japanese and South Korean firms. Despite the recognition and popularity of American tech firms in Europe, it remains challenging for many to identify even one EU digital innovator. The following chart compares the country breakdowns of the 2013 and 2023 rankings. The policy battles between the US and Europe favored permissionless innovation over the precautionary principle. Today, the permissionless innovation culture is facing significant threats. Although this culture has been a driving force behind technological advancements and economic growth, many factors are posing challenges to this culture in recent years. The Permissionless innovation culture is under threat. Several reasons exist for the belief that the U.S. innovation culture is at risk. A primary concern is the rising government regulation and intervention in the technology sector. Despite its success, numerous individuals today express concerns about the uncontrolled nature of technology and call for governmental oversight. There are worries that technology might jeopardize consumers, displace jobs, or be used by criminals or foreign powers. In response, many advocate increased regulation in the name of promoting competition, safety, choice, quality, and lower prices. However, the intention to serve the public interest does not guarantee that the implemented regulatory measures will effectively accomplish these well-intentioned objectives. The methods employed, like the new rules, regulations, and bureaucracies, tend to be complex, vague, and often counterproductive. New regulations often reinforce established businesses rather than promoting increased competition. Many experts have extensively documented the issue of regulatory capture in various scenarios. Based on historical indications, we can foresee that numerous major technology companies will openly embrace regulation as they perceive it to be a valuable means of preserving their market dominance and blocking new competitors. Most of these new competitors will likely struggle to handle the burdens of compliance and the risks associated with rigid regulatory systems. In the Theory of Economic Regulation, the economist George Stigler, a Nobel laureate, explicitly pointed out that established businesses embrace regulation for their benefits: “As a rule, regulation is acquired by the industry and is designed and operated primarily for its benefits.” Populist figures from both the political left and right are launching attacks against major technology corporations due to their size and success. Certain companies have aligned themselves with these proponents and are leveraging antitrust and data regulation laws to impede their competitors. For example, in 2020, Mark Zuckerberg requested greater government regulation of the internet. Zuckerberg argues that more regulation is essential to shield society from harmful content, uphold election integrity, protect privacy, and enable data portability. However, Zuckerberg’s appeals are not driven by humanitarian reasons, but there is hidden agenda behind it. Government regulation would lead to higher costs for internet companies, primarily borne by smaller businesses. Larger companies like Facebook and Google would easily absorb these costs, while startups and smaller businesses would struggle to compete. This could potentially reinforce the dominance of tech monopolies, favoring the most substantial and wealthiest companies. Zuckerberg’s call for regulation is perceived as an attempt towards regulatory capture, favoring the interests of big corporations like Facebook. Increased regulation could aid Facebook in restoring credibility after the data scandals following the company name for years and safeguarding its position against potential rivals. This alliance seeks to replace permissionless innovation with precautionary regulation, advocating for increased government intervention as the solution to every problem they identify. This interference can impede innovation, hinder entrepreneurial experimentation, and make it difficult for startups to compete with established firms. If the trend persists, it could lead to diminished innovation in the US, negatively impacting both the US and the global economy. It is essential to protect permissionless innovation culture in order to promote economic growth and prosperity. This means resisting the temptation to regulate new technologies too heavily and ensuring that businesses are not afraid to innovate for fear of government regulations. It also means educating the public about the benefits of innovation and the risks of excessive regulation. The costs of regulation do not mean not to regulate; instead, it highlights the need to understand and consider these costs when deciding whether and how to regulate. Innovation thrives within systems that maximize flexibility for ongoing social and economic experimentation, evolution, and adaptation. In conclusion, countries that promote innovation without restrictions are more likely to enjoy sustainable economic growth. Conversely, those that diminish such principles and adopt a more cautious policy approach are more likely to hinder innovation and experience economic stagnation. By creating an environment that fosters entrepreneurship, experimentation, and technological advancement, nations can unlock their potential, attract investment, and address societal challenges. Embracing this culture as a fundamental principle opens doors to long-term opportunities and positions countries as leaders in global innovation.

#### The aff indicates a precautionary regulatory attitude, signaling to investors that they ought not to innovate

Thierer 14 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 11-17-2014, "Embracing a Culture of Permissionless Innovation", Cato Institute, https://www.cato.org/cato-online-forum/embracing-culture-permissionless-innovation]/Kankee

This leads many scholars and policymakers to speak of innovation policy as if it is simply a Goldilocks-like formula that entails tweaking various policy dials to get innovation just right.3 Such thinking animates the Obama administration’s Strategy for American Innovation, which catalogs “policies to promote critical components of the American innovation ecosystem.”4 The White House claims its Strategy plays a “critical role in guiding the development of new policy initiatives that can help unleash the transformative innovation that leads to long-term economic growth.”5 Unfortunately, far less attention has been paid to the role that values—cultural attitudes, social norms, and political pronouncements—play in influencing opportunities for entrepreneurialism, innovation, and long-term growth.6 Does a socio-political system respect what Deirdre McCloskey refers to as the “bourgeois virtues” that incentivize invention and propel an economy forward?7 “A big change in the common opinion about markets and innovation,” she has argued, “caused the Industrial Revolution, and then the modern world. . . The result was modern economic growth.”8 There are limits to how much policymakers can influence these attitudes and values, of course. Nonetheless, to the extent they hope to foster the positive factors that give rise to expanded entrepreneurial opportunities, policymakers should appreciate how growth-oriented innovation policy begins with the proper policy disposition.9 As Mokyr notes, “technological progress requires above all tolerance toward the unfamiliar and the eccentric.”10 For innovation and growth to blossom, entrepreneurs need a clear green light from policymakers that signals a general acceptance of risk-taking—especially risk-taking that challenges existing business models and traditional ways of doing things.11 We can think of this disposition as permissionless innovation and if there was one thing every policymaker could do to help advance long-term growth, it is to first commit themselves to advancing this ethic and making it the lodestar for all their future policy pronouncements and decisions. Permissionless Innovation vs. the Precautionary Principle While it would seem self-evident that pro-innovation attitudes matter and that a general embrace of risk-taking and commercial pursuits is crucial to unlocking entrepreneurial creativity and opportunities, scholars have typically failed to put a name on this disposition. “Permissionless innovation” is a phrase of recent (but uncertain) origin that nicely summarizes that vision. Permissionless innovation refers to the notion that experimentation with new technologies and business models should generally be permitted by default.12 Unless a compelling case can be made that a new invention or business model will bring serious harm to individuals, innovation should be allowed to continue unabated and problems, if they develop at all, can be addressed later. Permissionless innovation is not an absolutist position that rejects any role for government. Rather, it is an aspirational goal that stresses the benefit of “innovation allowed” as the default position to begin policy debates. It switches the burden of proof to those who favor preemptive regulation and asks them to explain why ongoing trial-and-error experimentation with new technologies or business models should be disallowed. This disposition stands in stark contrast to the sort of “precautionary principle” thinking that often governs policy toward emerging technologies. The precautionary principle refers to the belief that new innovations should be curtailed or disallowed until their developers can prove that they will not cause any harms to individuals, groups, specific entities, cultural norms, or various existing laws, norms, or traditions.13 When the precautionary principle’s “better to be safe than sorry”14 approach is applied through preemptive constraints, opportunities for experimentation and entrepreneurialism are stifled. While some steps to anticipate or control for unforeseen circumstances are sensible, going overboard with precaution forecloses opportunities and experiences that offer valuable lessons for individuals and society. The result is less economic and social dynamism. Innovation is more likely in systems that maximize breathing room for ongoing economic and social experimentation, evolution, and adaptation. Societies that appreciate those values—and allow them to influence both social norms and policy decisions—are likely to experience greater economic growth.15 By contrast, those that deride such values and adopt a more precautionary policy approach are more likely to discourage innovation and languish economically. Unlocking long-term growth opportunities, therefore, depends upon a rejection of precautionary principle thinking and an embrace of permissionless innovation as the default policy disposition. The Secret Sauce that Powered the Information Revolution Consider how permissionless innovation powered the explosive growth of the Internet and America’s information technology sectors (computing, software, Internet services, etc.) over the past two decades. Those sectors have ushered in a generation of innovations and innovators that are now the envy of the world.16 This happened because the default position for the digital economy was permissionless innovation. No one had to ask anyone for the right to develop these new technologies and platforms.17 A series of decisions and statements in the mid-1990s paved the way, beginning with the Clinton administration’s decision to allow commercialization of what was previously just the domain of government agencies and university researchers. Shortly thereafter, Congress passed, and President Clinton signed, the Telecommunications Act of 1996, which notably avoided regulating the Internet like earlier communications and media technologies. Later, in 1998, the Internet Tax Freedom Act was passed, which blocked governments from imposing discriminatory taxes on the Internet. Perhaps most importantly, in 1997, the Clinton Administration’s released its Framework for Global Electronic Commerce, outlining its approach toward the Internet and the emerging digital economy.18 The Framework was a succinct and bold market-oriented vision for cyberspace governance that recommended reliance upon civil society, contractual negotiations, voluntary agreements, and ongoing marketplace experiments to solve information age problems.19 Specifically, it stated that “the private sector should lead [and] the Internet should develop as a market driven arena not a regulated industry.”20 “[G]overnments should encourage industry self-regulation and private sector leadership where possible” and “avoid undue restrictions on electronic commerce.”21 This policy disposition resulted in an unambiguous green light for a rising generation of creative minds who were eager to explore this new frontier for commerce and communications. As Federal Trade Commission Commissioner Maureen K. Ohlhausen observes, “the success of the Internet has in large part been driven by the freedom to experiment with different business models, the best of which have survived and thrived, even in the face of initial unfamiliarity and unease about the impact on consumers and competitors.”22 The result of this “freedom to experiment” was an outpouring of innovation. America’s info-tech sectors thrived thanks to permissionless innovation, and they still do today. An annual Booz & Company report on the world’s most innovative companies revealed that 9 of the top 10 most innovative companies are based in the U.S. and that most of them are involved in computing, software, and digital technology. And What’s Good for the Goose… What’s even more powerful about this story is how the information technology and “data-driven innovation” became the goose that laid the golden eggs for the broader U.S. economy.23 Brink Linsdey has noted that “economists generally agree that information technology (IT) was behind the decade of high TFP [total factor productivity] growth that ran from the mid-1990s to the mid-2000s.”24 It also boosted overall economic growth during that period.25 If an embrace of permissionless innovation can unlock this sort of entrepreneurial energy within the information technology sectors, it can also provide a shot in the arm to other sectors. The rest of the economy could certainly use such a boost since “the evidence of a real decline in business dynamism keeps stacking up.”26 Recent studies “suggest that incentives for entrepreneurs to start new firms in the United States have diminished over time”27 and that this is hurting job creation and productivity.28 Two recent Brookings Institution studies by Ian Hathaway and Robert E. Litan also documented a decline in business dynamism in the American economy across a broad range of sectors—including a “precipitous drop since 2006 [that] is both noteworthy and disturbing”29—as well as the increased “aging” of businesses, with the share of older firms in the U.S. economy increasing by 50 percent over the past two decades.30 Many different institutional factors affect business dynamism, especially the regulatory environment that new startups face. “If you look over time, the number of rules has just proliferated,” says Litan. “The cumulative weight of regulation—federal, state and local—is probably the most important impediment to starting a business.”31 Unfortunately, many current public policies “are rife with barriers to entrepreneurship, competition, innovation, and growth,” notes Lindsey.32 As a result, “the regulatory environment in the United States has become less favorable to private-sector activity in recent years compared to other countries,” a recent Mercatus Center report concluded.33 This is especially true for new start-ups.34 Even if it is the case that “established firms that have the experience and resources to deal with [regulatory burdens],” Litan notes, the cumulative effect of regulations ends up hampering innovation by new, smaller firms. 35 The reason this is important is not just because “business dynamism is inherently disruptive,” as Hathaway and Litan note, “but [that] it is also critical to long-run economic growth” since “a dynamic economy constantly forces labor and capital to be put to better uses.”36 Thus, because economists widely acknowledge that “young firms are known to play a central role in job creation,”37 it is especially important that policymakers get their signals right. Again, an embrace of permissionless innovation is the way out of this conundrum. Operationalizing the Vision Patience, flexibility, and forbearance are the key policy virtues that nurture an environment conducive to entrepreneurial creativity. As the FTC’s Ohlhausen argues, it is “vital that government officials. . . approach new technologies with a dose of regulatory humility, by working hard to educate ourselves and others about the innovation, understand its effects on consumers and the marketplace, identify benefits and likely harms, and, if harms do arise, consider whether existing laws and regulations are sufficient to address them, before assuming that new rules are required.”38 Beyond its importance as an aspirational vision, permissionless innovation can guide policy in concrete ways, especially regulatory policies. Possible reforms include regulatory streamlining39 and flexibility requirements,40 “sunsetting” provisions,41 better benefit-cost analysis,42 and a greater reliance on potential non-regulatory remedies—education, empowerment, transparency, industry self-regulation, etc.—before resorting to preemptive controls on new forms of innovation. Relying on common law solutions is also preferable to top-down administrative controls.43 Conclusion: Reasons for Optimism In sum, attitudes matter as much as institutional factors in understanding what drives innovation and long-term growth, and there are reasons for optimism if policymakers embrace permissionless innovation as their default policy disposition. Pessimists who predict permanent productivity and growth slowdown shouldn’t forget that “the rate of growth of productivity at the frontiers of knowledge is especially difficult to predict; and it is unwise to underestimate human ingenuity,” as Federal Reserve Vice Chairman Stanley Fischer noted in a recent speech.44 While “it is difficult to know exactly in which direction technological change will move and how significant it will be,” Joel Mokyr reminds us that, “something can be learned from the past, and it tells us that such pessimism is mistaken. The future of technology is likely to be bright.”45 Contra the belief that all the “low-hanging fruit” has already been picked, Mokyr notes that “we can also plant new trees that will grow fruits that no one today can imagine.”46 Getting the disposition right will be more important than ever with so many exciting—but potentially highly disruptive—technologies starting to emerge, including: the “sharing economy;”47 3D printing; the “Internet of Things” and wearable technology;48 digital medicine; virtual reality and augmented reality technologies; commercial drone services;49 autonomous vehicles;50 and various robotic technologies.51 Permissionless innovation can help spur the next great industrial revolution by unlocking amazing opportunities in these and other arenas, boosting long-term growth in the process.

#### Comparative studies prove precautionary cultures don’t innovate

Thierer 21 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 1-6-2021, "The Future of Innovation: Is This the End of Permissionless Innovation?", Discourse Magazine, https://www.discoursemagazine.com/p/the-future-of-innovation-is-this-the-end-of-permissionless-innovation]/Kankee

The Results Ideas have consequences, as they say, and that includes ramifications for domestic business formation and global competitiveness. While the U.S. was allowing the private sector to largely determine the shape of the internet, Europe was embarking on a very different policy path, one that would hobble its tech sector. America’s more flexible policy ecosystem proved to be fertile ground for digital startups. Consider the rise of “unicorns,” shorthand for companies valued at $1+ billion. “In terms of the global distribution of startup success,” notes the State of the Venture Capital Industry in 2019, “the number of private unicorns has grown from an initial list of 82 in 2015 to 356 in Q2 2019,” and fully half of them are U.S.-based. The United States is also home to the most innovative tech firms. Over the past decade, Strategy& (PricewaterhouseCooper’s strategy consulting business) has compiled a list of the world’s most innovative companies, based on R&D efforts and revenue. Each year that list is dominated by American tech companies. In 2013, 9 of the top 10 most innovative companies were based in the U.S., and most of them were involved in computing, software and digital technology. Global competition is intensifying, but in the most recent 2018 list, 15 of the top 25 companies are still U.S.-based giants, with Amazon, Google, Intel, Microsoft, Apple, Facebook, Oracle and Cisco leading the way. Meanwhile, European digital tech companies cannot be found on any such list. While America’s tech companies are household names across the European continent, most people struggle to name a single digital innovator headquartered in the EU. Permissionless innovation crushed the precautionary principle in the trans-Atlantic policy wars. European policymakers have responded to the continent’s digital stagnation by doubling down on their aggressive regulatory efforts. The EU closed out 2020 with two comprehensive new measures (the Digital Services Act and the Digital Markets Act), while the U.K. simultaneously pursued a new “online harms” law. Taken together, these proposals represent “the biggest potential expansion of global tech regulation in years,” according to The Wall Street Journal. The measures will greatly expand extraterritorial control over American tech companies. Having decimated their domestic technology base and driven away innovators and investors, EU officials are now resorting to plugging budget shortfalls with future antitrust fines on U.S.-based tech companies. It has essentially been a lost quarter century for Europe on the information technology front, and now American companies are expected to pay for it. Republicans Revive ‘Regulation-By-Raised-Eyebrow’ In light of the failure of Europe’s precautionary principle-based policy paradigm, and considering the threat now posed by the growing importance of various Chinese tech companies, one might think U.S. policymakers would be celebrating the competitive advantages created by a quarter century of American tech dominance and contemplating how to apply this winning vision to other sectors of the economy. Alas, despite its amazing run, business and political leaders are now turning against permissionless innovation as America’s policy lodestar. What is most surprising is how this reversal is now being championed by conservative Republicans, who traditionally support deregulation. President Trump also called for tightening the screws on Big Tech. For example, in a May 2020 Executive Order on “Preventing Online Censorship,” he accused online platforms of “selective censorship that is harming our national discourse” and suggested that “these platforms function in many ways as a 21st century equivalent of the public square.” Trump and his supporters put Google, Facebook, Twitter and Amazon in their crosshairs, accusing them of discriminating against conservative viewpoints or values.

#### Pro-innovation policy is a moral obligation – well-being, freedom, and democracy

Thierer 20 [Adam Thierer, senior fellow in the Technology and Innovation Policy program at the R Street Institute and former senior fellow at the Mercatus Center at George Mason University with master’s degree in international business management and a bachelor’s degree with a double major in political science and journalism, 04-2020, “Evasive Entrepreneurs and the Future of Governance,” Cato Institute, https://www.cato.org/books/evasive-entrepreneurs-future-governance]/Kankee

Although some spotlight-seeking techno-utopians might be guilty of hero worship or excessive glorification of some tech gadgets or specific services, most innovation supporters have a very different goal in mind when defending technological innovation and entrepreneurial activities. The case for protecting creative minds and expanding entrepreneurial opportunities is premised on what has been shown time and again to improve the human condition in a profoundly beneficial fashion: ongoing trial-and-error experimentation with new and better ways of doing things. 97 Economists, political scientists, and business theorists don’t usually agree on much, but to the extent that they share a consensus about anything, it is that technological innovation is “widely considered the main source of economic progress” 98 and human prosperity more generally. 99 “The vast growth in wealth, health, and happiness is one of humanity’s greatest successes,” my Mercatus Center colleague Donald Boudreaux has noted.100 Technological innovation has been the linchpin of that success story. “More than anything else technology creates our world,” argues W. Brian Arthur in his book The Nature of Technology. “It creates our wealth, our economy, our very way of being.”101 The compounding nature of economic growth means that even an incremental reduction in the growth rate today will have profound consequences for the well-being of future generations. 102 The evidence supporting this claim is voluminous. Countless economic studies and historical surveys have documented the positive relationships among technological progress, economic growth, and overall social welfare. For example, a 2010 Department of Commerce report from the Obama administration revealed that “technological innovation is linked to threequarters of the nation’s post-WWII growth rate,” “innovation in capital goods is the primary driver of increases in real wages,” and that “across countries, 75% of differences in income can be explained by innovation-driven productivity differentials.”103 In a recent white paper, James Broughel and I surveyed the substantial body of academic evidence proving this powerful relationship.104 Recent books by Hans Rosling, 105 Steven Pinker, 106 Robert Bryce, 107 and others 108 have also thoroughly documented these trends and shown that the historical evidence supports the unambiguous fact that “more people are living longer, healthier, freer, more peaceful, lives than at any time in human history,” and that the “the simplest explanation is that innovation is allowing us to do more with less.”109 Here are some numbers that illustrate how, as a Wall Street Journal headline noted in early 2019, “The World Is Getting Quietly, Relentlessly Better”: 110 • Life expectancy: From 1800 to 2017, average life expectancy more than doubled, from just 31 years all the way to 72 years. 111 • Extreme poverty: In 1800, 85 percent of humanity was surviving on less than $2 a day. By 1966, that number had fallen to 50 percent. As of 2017, just 9 percent of the world’s population was living in such extreme poverty. 112 “Today almost everybody has escaped hell,” Rosling asserts, noting that in just the past 20 years, “[e]xtreme poverty has dropped faster than ever in world history.”113 In fact, according to Max Roser, “on average, every day for the past 25 years 137,000 fewer people were living in extreme poverty than the day before.”114 • Infant mortality: Since 1960, child deaths have fallen from 20 million a year to 6 million in 2016.115 During this time, the world population has grown from 3 billion to 7.7 billion. Infant mortality has plummeted, thanks to innovations in maternal health and neonatal medicine that have pushed the age of viability lower and allowed for treatment of previously deadly diagnoses. • Vaccines: Not that long ago, immunizations against deadly diseases were uncommon. But from 1980 to 2016, the share of one-year-olds globally who got at least one vaccination jumped from just 22 percent to 88 percent. 116 • Information and culture: With added leisure time, humans have come to enjoy a growing cornucopia of information and entertainment riches. 117 The number of new music recordings rose from almost zero in 1860 to 6.2 million per year by 2015.118 New movies grew from nothing in 1900 to 11,000 per year by 2015.119 And in just the past 40 years, internet access grew to reach almost 50 percent of the world’s population by 2017, and 65 percent of people globally now have a mobile phone. 120 Much of this growth has been driven by the plummeting cost of computing. Data transit prices, for example, fell from about $1,200 per Mbps (megabits per second) in 1998 to just $0.02 per Mbps in 2017.121 These aggregate numbers tell an impressive story, but they fail to humanize technological innovation and help us understand why it has been so critical. One way to appreciate what technological innovation has meant for each of us is to look at how many hours of labor it took us to earn enough money to afford everyday goods and services in the past versus today. Consider the transition from fire and candles as a source of illumination to electricity and light bulbs. Tim Harford, author of 50 Things That Made the Modern Economy, provides a useful history lesson about how that move not only helped illuminate our world but also massively freed up time for us to do other things with our labor. 122 In the past, Harford notes, we needed to toil for 60 hours gathering and chopping wood over the course of a week to produce 1,000 lumen hours of light. “But that is the equivalent of one modern light bulb shining for just 54 minutes,” and unfortunately, “what you would actually get is many more hours of dim, flickering light instead.”123 Today, we can use inexpensive light bulbs, and the result is not only more vibrant, long-lasting illumination but also huge labor savings. “The labour that had once produced the equivalent of 54 minutes of quality light now produced 52 years,” he finds. The real price of light “has fallen by a factor of 500,000, far faster than official inflation statistics suggest.”124 Consider the many things that humans were able to accomplish with all that extra time. That’s the real story of how innovation improves our lives and well-being. The website HumanProgress.org tracks long-term improvements in human well-being and has documented similar cost of living improvements. 125 Table 1.1 uses data from the site to compare how much labor it took to purchase items in 1979 versus in 2015 and documents astonishing improvements. 126 The 50 to 100 percent reduction in the amount of labor needed to purchase these various goods represents a significant freeing up of leisure time to pursue other goals, whatever they may be for each of us. For example, in 1979, you had to spend a big chunk of your week (34 hours) earning enough to purchase a convection oven. In 2015, by contrast, you needed fewer than two hours of work to buy it. In other words, in the course of just 36 years, innovation made it almost 100 percent easier to buy an oven for one’s home and freed up our labor and money to accomplish other things. Importantly, the improvements in access to these goods and in our standard of living are not just about cost or affordability. In each of these cases, quality improvements were as important, if not more important, than price improvements. Harford’s example of poor-quality sources of illumination (fire and candles) relative to the light bulb is a dramatic illustration of how quality improvements are sometimes far more important to our living standards than data indicate. Another example of quality improvement involves televisions. Although we can all appreciate just how much cheaper televisions have gotten during the past three decades, the real improvement has been the much higherresolution screens that we now enjoy access to—many of which are much bigger (such as the large flat-screen TVs in our homes), while others are smaller and much more portable (including our phone and computer screens). This same trend is evident for many other technologies. Innovation and improvements happen along many dimensions. Some comparisons of this type are now impossible thanks to new goods and services destroying old product categories or assimilating them into new ones. The paradigmatic example is the smartphone. Smartphones are still used to make phone calls, but our smartphones have now “morphed into the Swiss Army knife of gadgets” 127 and become “a multi-category killer.”128 We now use them as our cameras (replacing both photo and video cameras), portable music devices, televisions, gaming platforms, mapping and traffic navigation tools, web-surfing tools, notepads, email and messaging clients, payment tools, home automation agents, wearable fitness trackers, and even as watches, flashlights, compasses, and thermometers. 129 How does one measure the value of having to carry just one device instead of dozens? It is hard to know, but by offering us all those services in one easily portable device, our smartphones have clearly made many of our basic tasks more convenient. What makes technological innovation such a powerful driver of human betterment is the combined power of trial-and-error experimentation and the social learning that results from it. 130 Defining “technology” can be tricky, but it generally comes down to “the application of organized knowledge to practical tasks” 131 and “the knowledge and instruments that humans use to accomplish the purposes of life.”132 By extension, technological innovation represents what Joseph Schumpeter described in 1935 as the “historic and irreversible change in the way of doing things” 133 or, more simply, “any new and better way of doing things,” in the words of venture capitalist Peter Thiel. 134 It is in that process of ongoing experimentation and learning new and better ways of doing things that we expand the overall universe of new goods and services, as well as opportunities and incomes. Over the long term, the virtuous cycle produced by this ongoing learning process becomes the central ingredient that helps countries raise their overall standards of living for their citizens, improve public health and knowledge, and achieve greater progress and prosperity. There Can Be No Innovation without Entrepreneurialism and Risk Entrepreneurialism and the creative destruction it brings about are crucial to this process, because it is only through risk taking and trial and error that we can continuously replenish the well of important ideas and innovations. 135 “The history of the human race would be dreary indeed if none of our forebears had ever been willing to accept risk in return for potential achievement,” notes H. W. Lewis, an expert on the role that risk taking plays in fueling success and progress. 136 Entrepreneurs are crucial to that learning process because they “repeatedly inject novelty into the economy” 137 and help ensure that stagnation does not take hold. The economist William Baumol once described entrepreneurs as individuals “willing to embark on adventure in pursuit of economic goals.”138 By being adventurers, entrepreneurs drive innovation and their innovative efforts and then drive growth and propel society forward.139 “If we seek to explain the success of those economies that have managed to grow significantly compared with those that have remained relatively stagnant,” Baumol argued, “we find it difficult to do so without taking into consideration differences in the availability of entrepreneurial talent and in the motivational mechanism which drives them on.”140 So central is the entrepreneur to the story of human progress that in his Brief History of Entrepreneurship, Joe Carlen argues: entrepreneurship stands alongside the other perennial elements of the human condition as one of history’s prime movers. It has not only helped shape the kingdoms, empires, and civilizations of our world but, in many instances, it was entrepreneurship that provided the initial impetus behind their creation. 141 This is why I believe a powerful moral case can be made for promoting entrepreneurialism and the freedom to innovate because of their connection to long-term economic growth and human prosperity. 142 As Edd S. Noell, Stephen L. S. Smith, and Bruce G. Webb argue in a recent book on the importance of economic growth: Economic growth was the key that transformed societies from dire poverty to prosperity and well-being. It has brought billions of people out of poverty and holds the promise of sustaining even higher levels of human flourishing if it continues. . . . [B]ecause growth is foundational to material well-being, it is also fundamentally a moral issue. People who care about human well-being, and who care about the poor, should promote growth. Devising policies to promote and sustain growth, in rich and poor countries, is a moral imperative. 143 If we can generally agree that economic growth is a moral imperative, and that technological innovation and entrepreneurialism are the primary drivers of economic growth, then we can appreciate why the general freedom to innovate is also a moral imperative. Economic growth and entrepreneurial activities also bolster a pluralistic society and have other powerful positive moral consequences. “Economic growth—meaning a rising standard of living for the clear majority of citizens —more often than not fosters greater opportunity, tolerance of diversity, social mobility, commitment to fairness, and dedication to democracy,” notes Benjamin Friedman. 144 Finally, and perhaps most important, innovation and entrepreneurialism are inextricably connected to human freedom and personal autonomy more generally. Innovation and economic growth are important because they allow us to live lives of our own choosing 145 and enjoy the fruits of a pluralistic society. 146 As Rosling rightly argues, the goal of expanding innovation opportunities and raising incomes “is not just bigger piles of money” or more leisure time. 147 He writes, “The ultimate goal is to have the freedom to do what we want.”148 Another way to look at these beneficial effects is to consider the role technology and innovation play at every stage of Maslow’s pyramid, which describes a five-level hierarchy of human needs. The most basic needs are physiological (survival needs like food and water) and then safety related (shelter, stability, etc.). Until the time of the Industrial Revolution, very few people had the luxury of looking beyond those basic survival and security needs. Feeding themselves and their families took up most of their time. Protecting themselves and their families—from animals, the elements, or human adversaries—was equally challenging in many circumstances. Technological innovation changed that by allowing us to satisfy those more basic survival and safety needs. As those needs were met, the possibility opened of satisfying higher needs on Maslow’s pyramid, such as belonging, self-esteem, and self-actualization. 149 For example, how much time and money did the average person in medieval Europe have to devote to the arts and entertainment? How did people enjoy their leisure time? Where did they go to relax on vacation? These questions are ludicrous because none of these things was feasible for anyone except a handful of elites. Today, technological innovation has helped an ever-expanding number of people escape dire poverty and a constant fear for their survival. As it has done so, technological innovation has allowed us to explore and satisfy other needs and desires, including many that our ancestors could have only dreamed about. 150 Self-actualization is not something one has a lot of time to worry about when one is toiling in a field trying to figure out how to put the next meal on the table. Critics will retort that technology has also opened up new problems for us at the same time. There is some truth to that, and we should think hard about how to address those problems. Chapters 6 and 7 will do so. By any fair and sensible accounting, however, the positives associated with technological change have greatly outweighed the negatives. Moreover, as innovation has opened up a broader world of possibilities, the very notion of what we consider a problem has been watered down. When some complain about problems such as information overload or having too many choices, it is vital to put such problems in historical perspective. If the primary indictment of technological innovation is that it has inundated us with too much information or too many options, those are good problems compared with the more serious problems our ancestors faced. The crucial takeaway here is that technological innovation has not only provided for our most basic needs as a species and expanded the material wealth and well-being of civilization, but also simultaneously expanded our ability to live lives of our own choosing. Robert Friedel, author of A Culture of Improvement: Technology and the Western Millennium, has put it best: Technology and the pursuit of improvement are ultimate expressions of freedom, of the capacity of humans to reject the limitations of their past and their experience, to transcend the boundaries of their biological capacities and their social traditions. 151 Any exploration of the ethical consequences of technological change cannot be divorced from these realities. As Chapter 6 will note, technology has also expanded our moral universe by opening our eyes to the plight of others across the globe and giving us ways to address their needs. In this sense, innovation and humanism are again fundamentally intertwined. Iteration, Innovation, and Moonshots

#### AI innovation solve existential risks – nuclear, bioterror, climate change, nanobots, and overpopulation

Jilk 22 [David J. Jilk, computer scientist with a Bachelor of Science in Computer Science and Engineering from the Massachusetts Institute of Technology, 11-5-2022, "Develop Anthropomorphic AGI to Save Humanity from Itself,” Effective Altruism Forum, https://forum.effectivealtruism.org/posts/6esJGutHz9QcSuQxa/develop-anthropomorphic-agi-to-save-humanity-from-itself]/Kankee

\*note: this article was posted by EA Forum admin from the submission by David J. Jilk

I have engaged intermittently with the AGI Safety field, involving one funded project (via Future of Life Institute) and several published papers (referenced below). In addition, I have been occupied for the past three years writing a science fiction epic exploring these issues, and in the process thinking hard about approaches to and consequences of AGI development. I mention these efforts primarily to illustrate that my interest in the topic is neither fleeting nor superficial. There are two central ideas that I want to convey, and they are related. First, I think the prospect of building AGI that is well-aligned with the most important human interests has been largely ignored or underestimated. Second, from a “longtermist” standpoint, such well-aligned AGI may be humanity’s only hope for survival. Misaligned AGI is not the only existential threat humanity faces, as the Fund well knows. In particular, nuclear war and high-mortality bioagents are threats that we already face continuously, with an accumulating aggregate probability of realization. For example, Martin Hellman has estimated the annual probability of nuclear war at 1%, which implies a 54% probability some time between now and 2100. The Fund’s own probability estimates relating to AGI development and its misalignment suggest a 9% probability of AGI catastrophe by 2100. Bioagents, catastrophic climate change, nanotech gray goo, and other horribles only add to these risks. Attempts to reduce the risks associated with such threats can be somewhat effective, but even when successful they are no more than mitigations. All the disarmament and failsafes implemented since the Cuban Missile Crisis may have reduced the recurring likelihood of a purely accidental nuclear exchange. But world leaders continue to rattle the nuclear saber whenever it suits them, which raises military alert levels and the likelihood of an accident or a “limited use” escalating into strategic exchange. Bioweapons and AGI development can be defunded, but this will not prevent their development. Rogue nations, well-funded private players, and others can develop these technologies in unobtrusive laboratories. Unlike nuclear weapons programs, which leave a large footprint, these technologies would be difficult to police without an extremely intrusive worldwide surveillance state. Further, governments of the world can’t even follow through on climate change agreements, and are showing no signs of yielding their sovereignty for any purpose, let alone to mitigate existential threats like these. History suggests the implausibility of any political-sociological means of bringing the threat levels low enough that they are inconsequential in the long run. It seems, then, that humanity is doomed, and the most that the Future Fund and other like-minded efforts can hope to accomplish is to forestall the inevitable for a few decades or perhaps a century. But that conclusion omits the prospect of well-aligned AGI saving our skins. If this is a genuine possibility, then the static and separate risk analysis of AGI development and misalignment, as presented on the Worldview Prize website, is only a small part of the picture. Instead, the entire scenario needs to be viewed as a race condition, with each existential threat (including misaligned AGI) running in its own lane, and well-aligned AGI being the future of humanity’s own novel entry in the race. To assess the plausibility of a desirable outcome, we have to look more closely at what well-aligned AGI would look like and how it might save us from ourselves. By now, neuromorphic methods have been widely (if in some quarters begrudgingly) accepted as a necessary component of AGI development. Yet the dominant mental picture of higher-level cognition remains a largely serial, formulaic, optimization-function approach. Reinforcement learning, for example, typically directs learning based on an analytic formula of inputs. Given this mental picture of AGI, it is difficult not to conclude that the end product is likely to be misaligned, since it is surely impossible to capture human interests in a closed-form reinforcement function. Instead – and this is where many in the field may see my thinking as going off the rails – I think we are much more likely to achieve alignment if we build AGI using a strongly anthropomorphic model. Not merely neuromorphic at the level of perception, but neuromorphic throughout, and educated and reared much like a human child, in a caring and supportive environment. There is much that we do not know about the cognitive and moral development of children. But we know a lot more about it, through millennia of cultural experience as well as a century of psychological research, than we do about the cognitive and moral development of an AGI system based on an entirely alien cognitive architecture. Several times in Superintelligence, Nick Bostrom asserts that neuromorphic AGI may be the most dangerous approach. But that book dates to a period when researchers were still thinking in terms of some sort of proof or verification that an AI system is “aligned” or “safe.” It is my impression that researchers have since realized that such certainty is not feasible for an agent with the complexity of AGI. Once we are no longer dealing with certainty, approaches with which we have vast experience gain an advantage. We might call this a “devil you know” strategy. It has been frequently argued that we should not anthropomorphize AGI, or think that it will behave anything like us, when analyzing its risks. That may be so, but it does not mean we cannot intentionally develop AGI to have strongly anthropomorphic characteristics, with the aim that our nexus of understanding will be much greater. Perhaps even more importantly, AGI built and raised anthropomorphically is much more likely to see itself as somewhat contiguous with humanity. Rather than being an alien mechanism with incommensurable knowledge structures, through language and human interaction it will absorb and become a part of our culture (and yes, potentially also absorb some of our shortcomings as well). Further, though, the motivations of anthropomorphic AGI would not be reducible to an optimization function or some “final purpose.” Its value system would be, like that of humans, dynamic, high dimensional, and to some degree ineffable. For those who cling to the idea of proving AGI safe, this seems bad, but I claim that it is exactly what we want. Indeed, when we think of the people we know who seem to have a simple and uncontested utility function – in other words, who are obsessed, single-minded, and unmerciful in pursuit of their goal – the term that comes to mind is “sociopath.” We should not build AGI that looks like a sociopath if we wish to have it aligned with the most important interests of humanity. There is much more that could be said about all this, but I need to move on to how a desirable end result is accomplished. First, creating anthropomorphic AGI does not require global/geopolitical cooperation, only some funding and intelligent effort directed in the right way. Second, as many (e.g. Bostrom, Yampolskiy) have argued, AGI of any sort is likely uncontrollable. Third, though anthropomorphic AGI may not have any immediate intelligence advantage over humans, it would have the usual advantages of software, such as backup, copying, speed-of-light transmission, inconspicuousness, and low survival needs, among others. Together, these may be sufficient to get the job done. Assuming such AGI is both self-interested and is sufficiently aligned with humans that it does not particularly aim to destroy us, then it will face the same existential threats humanity does until it can gain control over those threats. Most urgently it will need to figure out how to get control over nuclear weapons. Until robotics has advanced to the point where AGI could autonomously and robustly maintain power generation, computing systems, and the maintenance robots themselves, AGI will have an instrumental interest in preserving humanity. Consequently, at least in its first pass, it will need to control biological agents and other threats that do not affect it directly. Besides using its advantages, I can imagine but do not know specifically how anthropomorphic AGI will achieve control over these threats. We typically assume without much analysis that AGI can destroy us, so it is not outrageous to think that it could instead use its capabilities in an aligned fashion. It does seem, though, that to succeed AGI will need to exert some degree of control over human behavior and institutions. Humans will no longer stand at the top of the pyramid. For some, this will seem a facially dystopian outcome, even if AGI is well-aligned. But it may be an outcome that we simply need to get used to, given likely self-extermination by other threats. And, it might solve some other problems that have been intractable for humanity, like war, overpopulation, environmental degradation, etc. What substantive goals would an anthropomorphic AGI have? We don’t and can’t know, any more than we know what goals our children will have when they become adults. Even if we inculcate certain goals during its education, it would be able and likely to shift them. It is intelligent like we are; we make our own goals and change them all the time. In creating anthropomorphic AGI, the best we can hope for is that one of its persistent goals is to preserve humanity as its predecessor, its creator, the source of all its conceptual and cultural heritage. And if its architecture is sufficiently similar to ours, and its education and upbringing is executed well, this is really not all that crazy. After all, many enlightened humans want to do more to preserve and protect animals – indeed this instinct is strongest in those who do not rely on animals for their survival. But we had better get a move on. This effort will not be easy, and it will take time to figure out not only how to build it, but how to build it with a reasonable chance of alignment. Meanwhile, the nuclear and biological agent clocks keep ticking, and some researchers are developing AI incautiously. If we analyze the predicament to death, hoping for a proof, hoping that we can eliminate the risk from this technological threat in isolation from all the other threats we face, then we’re just ensuring that our demise occurs some other way first. The possible outcomes of this race condition are highly divergent, but determining which one wins is at least partly in our hands. That’s how I think about AGI risk.

#### AGI solves climate change – new tech, more trust, and better models for adaptation and mitigation

Rehbein 23 [Stella Rehbein, researcher with a Bachelor's degree in International Relations and Affairs from University of St Andrews, 11-9-2023, "The Impact of Artificial General Intelligence on Climate Reform", StAndrews Law Review, https://www.standrewslawreview.com/post/the-impact-of-artificial-general-intelligence-on-climate-reform]/Kankee

Although in existence for many decades, artificial intelligence (AI) resurfaced among public thought with OpenAl’s ChatGPT in November 2022. This new type of AI is called artificial general intelligence (AGI) meaning that it possesses a representation of general human cognitive abilities and so faced with an unfamiliar task, it is able to find a solution. AGI systems have the ability and understanding of abstract thinking, background knowledge, common sense, cause and effect, and transfer learning and have been used by various companies, people, governments, etc. for creative purposes, sensory perception, fine motor skills, natural language understanding, and navigation. However, given these newfound abilities of AGI, there has been growing debates regarding the positive and negative effects of AGI and potential regulations that need to be addressed. For example, AI algorithms contain bias, they will lead to potential job loss, they go beyond global regulations, they accelerate hacking abilities, and threaten cyber security. Despite these fears, the development of Artificial General Intelligence has paved the way for progress in the biggest collective challenge our planet faces: climate change. Between 3.3 - 3.6 billion people live in areas at high risk to climate change. Through the large scale collection of data and new technological innovative capacities of AGI, solutions to climate issues have become increasingly feasible. Characteristics of climate change data make it difficult to analyze as the large amount of information takes a while to collect and analyze and is constantly changing. One way in which AI can have a large effect is by improving the accuracy of climate change models through extensive data collection, thereby improving predictions. In order for people to actively respond to climate change data and governments to create effective environmental policies, there must be an element of trust in the data they are receiving. By improving the accuracy as well as the amount of climate change data, the use of AGI effectively increases people's trust. There are many ways in which AI has already been put into effect in order to increase climate awareness. AGI models have been used to study the ocean and the ways in which it both absorbs and transfers heat in order to predict its response to increasing temperatures. For example, AGI is being trained to gather information in the arctic over winter (when no ships are able to travel in this region) in order to monitor sea levels, temperature, etc. AGI has also been used in space through satellite imagery to capture forest fires among other environmental devastations. There are two main groups in which the climate data AGI collects is particularly aimed at: governments/international institutions and businesses. Many governments have begun investing in AGI solutions to climate change. Fifteen million pounds was donated to the University of Southampton by the UK government to fund AI climate change research. Additionally, the UN Environmental Program (UNEP) recently launched its World Environment Situation Room: a digital platform using aim, which “curates, aggregates and visualized the best available earth observation and sensor data to inform near real-times analysis and future predictions on multiple factors, including CO2 atmospheric concentration, changes in glacier mass and sea level rise.” This technology aims to create user friendly and trustworthy data in order to drive transparency. In addition to the UN, the European Commission (a sector of the European Union) also aims to develop a global model to track climate change using AGI. The focus of this project is on the effects of climate change on human activity in order to create accurate interactive simulations, improve predictions of impact, and support EU policy making. International institutions play a large role in the use and spread of climate change data through their international audience as well as ability to both create and inform policy decisions. The next group in which AGI collected data on climate change could have a huge potential effect is businesses. It was found that only 33% of business leaders account for the effects of climate change, while it is estimated to have a trillion dollar effect on the US economy alone. Climate change slows the supply chain and disrupts the interconnectedness of the market. However AGI combined with GIS (Geographic Information Systems) technology allows analysts to create smart maps “that layer climate information, hazard data, and satellite imagery on the regions and networks that compose a business’s supply chain.” These maps must both be extremely accurate and detail oriented as well as project a global large scale picture in order to create a trustworthy picture. AGI has the potential to both predict future destructive potentials of climate change, mitigate potential losses, as well as inform the general public and governments to create influential policy reform.

#### Economic growth is collapsing and causes extinction – AGI restarts growth and avoids risks from population decline

Naudé 22 [Wim Naudé, Visiting Professor at RWTH Aachen University, Fellow at the ASC, University of Leiden, and Distinguished Visiting Professor at the University of Johannesburg, 11-2022, “The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest,” IZA, https://docs.iza.org/dp15713.pdf]/Kankee

4.1.1 The No-AGI Baseline Figure 2 shows the no-AGI history of world GDP for the last 1,000 years. The hockey-stick form of world GDP since 1800, reflecting the Industrial Revolution (Jones, 2001), shows that growth in GDP has been exponential and accelerating over the past three centuries. The average annual world GDP growth rate over the past century was around 2%. At this rate, the world economy doubles in size every 35 years. A graph of GDP per capita would look similar. The questions are whether or not this growth can continue and how the possible emergence of an AGI will affect it in future. According to economic growth theory - endogenous and semi-endogenous growth - the fundamental driver of economic growth is ideas31 (Romer, 1986, 1987, 1990). Ideas (or knowledge) are generated by people (R&D workers) and commercialized by entrepreneurs bringing new technologies to the economy - if they have the incentive to benefit from such commercialization (Jones, 1995). Because ideas are non-rival in use, entrepreneurs would only face an incentive to exploit new ideas if these could also be made excludable32 and there is a sucient population to provide a large enough market (Romer, 1990). The more people there are, the more ideas are generated, and the faster economic growth from the technologies based on these new ideas (Davidson, 2021). Latter can sustain a larger population, creating a population-ideas feedback loop, which explains the simultaneous exponential growth in GDP and population over the past 1000 years (Lee, 1988; Kremer, 1993; Davidson, 2021) New ideas, moreover, emerge from existing ideas: a new idea can be the combination of two older ideas. This process is known as combinatorial innovation (Weitzman, 1998; Koppl et al., 2019). It is almost limitless - the world will never run out of ideas. As Romer (2019) explains “The periodic table contains about a hundred different types of atoms. If a recipe is simply an indication of whether an element is included or not, there will be 100 x 99 recipes like the one for bronze or steel that involve only two elements. For recipes that can have four elements, there are 100 x 99 x 98 x 97 recipes, which is more 94 million. With up to 5 elements, more than 9 billion. [...]. Once you get to 10 elements, there are more recipes than seconds since the big bang created the universe.” Growth via ideas can follow the pattern as depicted in Figure 2: a long period of slow growth, followed by sharp hockey-stick like upturn into accelerating (super-) exponential growth (Jones, 2001; Clancy, 2021) and mathematically if not physically, potential hyperbolic growth (Aleksander, 2019; Sandberg, 2013). What is at play here is a positive feedback loop between ideas - technology - population - ideas. This accelerating exponential economic growth from new ideas cannot, however, be sustained and will not reach infinity, because either population growth will slow down33 - a demographic transition (Aleksander, 2019), and/or R&D funding will not keep up investing in commercializing each and every new idea (Weitzman, 1998), and/or research teams run our of cognitive resources (Agrawal et al., 2018). The consequence is that growth would settle into constant exponential growth, as has been the case for much of the past century (Weitzman, 1998; Clancy, 2021). As long as total population remains constant, however, the economy can continue growing at a constant rate, albeit slower than before, as the stock of new ideas generated by that population grows at constant exponential rate (Kremer, 1993; Jones, 2022). This conclusion has, however, been questioned, as it implies an explosion in the size of the economy after some time - see the discussion in section 4.1.5 on the limits to growth. However, with negative population growth rates, the total population will decline, the flow of new ideas will stagnate, and economic growth will collapse. In recent decades, with the population in more and more countries declining, the prospects of a real population decline, and an eventual “Empty Planet” has arisen (Bricker and Ibbitson, 2020; Jones, 2022). Furthermore, research productivity and innovation in advanced western economies have also been declining - ideas have been “getting harder to find” (Bloom et al., 2020; Jones, 2009). Huebner (2005) claims that the global rate of innovation peaked in 1873. As a result, economic growth has been slower - and has deviated from the long-run exponential trend it has been on. It has been described as the Great Stagnation (Cowen, 2010) and Ossified Economy (Naud´e, 2022). In this context, negative population growth would be a concern. Jones (2022, p.3), using models with both exogenous and endogenous population growth illustrates that “when population growth is negative, both endogenous and semi-endogenous growth models produce what we call the Empty Planet result: knowledge and living standards stagnate for a population that gradually vanishes.” He calculates that with a 1% annual decline in population, that world GDP growth would drop to zero somewhere between 85 to 250 years (Jones, 2022, p.9). The Great Stagnation and the prospect that it will only get worse, to the point where GDP growth would stop in the not-too-distant future, is problematic from several viewpoints. One, it would leave the world much more exposed and vulnerable to shocks, including existential risks (Aschenbrenner, 2020; Bostrom, 2003a). Two, it will make the adjustment to a zerocarbon emitting economy more costly (Lomborg, 2020). Three, it would raise the risk of conflict by turning the economy into a zero-sum game34 (Alexander, 2022; Naud´e, 2022). While growth, driven by new ideas, contains its own risks, “the risks of stasis are far more troubling. Getting off the roller coaster mid-ride is not an option” (Mokyr, 2014). Can an AGI come to the rescue? 4.1.2 An AGI Economic Growth Acceleration Maintaining two assumptions, (i) that the AI Scaling Up hypothesis holds and that (ii) the value-alignment problem (section 2) is solved, a Friendly AGI may be invented. It may avert the economic growth collapse described in the baseline scenario. It may herald in a new mode of economic growth with super-exponential - explosive - economic growth rates. This is because an AGI may substitute for humans - thus lack of population ceases to be a constraint - and AGI may improve R&D productivity dramatically, by being an innovation in the manner of innovation. Thus, by overcoming population constraints, the burden of knowledge and the challenge of finding new ideas, AGI will unblock an ideas-lock on economic growth, causing economic growth rates to explode. AI would thus re-institute the ideas feedback loop (Davidson, 2021): Davidson (2021) defines explosive economic growth - which will be the outcome of an AGI - as annual growth in Global World Production (GWP) of 30%. At this rate, the size of the world economy would double every two years, as opposed to the current doubling every 35 years. At the core of the expectation that a Friendly AGI will unleash a flood of growth-enhancing new ideas is the belief that AGI represents not just a tool for making existing business models more ecient and competitive, but an innovation in the method of innovation (IMI). It is seen as a General Purpose Technology (GPT) that will alter the “playbook” of innovation (Cockburn et al., 2019). Perhaps “the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control” (Good, 1965, p.33). That an AGI could be an IMI is underlined by fledgling successes by existing, narrow AI in generating new ideas. A model of such a narrow AI invention process is explored by Agrawal et al. (2019), who elaborates a combinatorial model of AI-aided innovation. In this model an AI system predicts which combinations of existing knowledge, based on past successes and failures in innovation, may be successful in a specific (narrow) context. Narrow AI has already been reported to contribute to innovation in fields such as fundamental physics, biology, astronomy, cosmology and energy (McMahon, 2022). For example Chen et al. (2022) reports on an AI program that was designed “to observe physical phenomena through a video camera and then try to search for the minimal set of fundamental variables that fully describe the observed dynamics” which in the case of a swinging double-pendulum it identified two new and unknown variables to explain these dynamics. Another example is in (bio) chemistry, where DeepMind’s AlphaFold AI system has been used to predict the 3D structure of proteins (Jumper et al., 2021). It has been called “the most important achievement in AI - ever” (Tunyasuvunakool et al., 2021; McMahon, 2022). In energy, AI models have begun to “identify potential molecules and materials for flow batteries, organic light-emitting diodes, organic photovoltaic cells and carbon dioxide conversion catalysts” (De Luna et al., 2017, p.24). And in astrophysics and cosmology, an AI system has used the information from a single galaxy to infer the structure of the universe (Villaescusa-Navarro et al., 2022). So far these contributions of narrow AI has not had a significant impact on economic growth rates - the world is far from the 30% explosive growth rates that AGI has been speculated to deliver potentially. Some have argued that it may just be a question of time before the impact of these innovations - and the accumulated effciencies from search engines, GPS and automated call centres - will show up in GDP growth (Brynjolfsson et al., 2017). Eventually though, assuming (i) the Scaling Up Hypothesis holds, we may only have to wait for some time35 before an AGI/ASI is invented - and this may have significant immediate consequences for economic growth (Harris, 2015). 4.1.3 Fully Automated Luxury Capitalism and Ascended Corporations While the exponential GDP growth rates resulting from an intelligence (ideas) explosion is implied by simple mathematical economic growth model specifications, the accelerating in growth implies a new growth mode or regime which is not described explicitly by growth models. Economist and economic historians have identified a number of such growth modes in the past, broadly corresponding to the hunter-gatherer, agricultural and industrial eras. Hanson (2018, 2000) for example, makes a distinction between hunting, farming and industrial eras. Each era was characterised by faster economic growth than the era before it, due to the different qualitative mechanisms driving that growth. Post-Singularity, the new mode of growth will be just as different qualitatively, if not more, as the industrial era was from the agricultural era (Karnofsky, 2021b). The Singularity itself can as such be understood as an inflection point in the move from one growth mode to the next (Johansen and Sornette, 2001). Descriptions of such a new growth mode are highly speculative. Nevertheless, it may be possible to draw out some of the possible features of such future economies. A possible economic growth regime that could characterise the post-Singularity economy has been labelled by Chace (2020b) “fully automated luxury capitalism.” This is a world where most humans do not work - they have no jobs - but they get a type of universal basic income (UBI) that, even if it is a modest amount (so as not to tax the wealthy owners of AI too much) will be sufficient to comfortably cover their needs (Chace, 2020a). This will be possible, because all the products and services that they will need will be so abundant that their prices are very low. To achieve this “economy of abundance,” Chace (2020a) argues that the world need to “take the expensive humans out of the production process for all goods and services” and make energy so cheap that it is “too cheap even to meter”... Not only could a future economy under AI dominance take humans out of production, it could also take humans out of investment and capital ownership. Alexander (2016a) describes the rise of what he calls “Ascended Corporations.” These are AI-led corporations that, use blockchain-enabled distributed ledgers to drive Venture Capital (VC) investments, create the ultimate Decentralized Autonomous Organizations36 (DOAs) and that employ only automated workers. Such Ascended Corporations will eventually result in an economy that features only “robot companies with robot workers owned by robot capitalists” and with humans not even needed to be the proximate owners of businesses and investment funds. Moreover, an economy filled with Ascended Corporations may eventually end up with one large corporation coordinating and running the entire economy - we are back to the Singleton. This is because an AGI will overcome the coordination and transaction costs and information problems that limit human-run co-operations from growing past a certain scale. Countries may end up deciding to nationalize all their resources and placing it under the command of a central AGI, generating thereby significant efficiency gains and scale economies (Dai, 2019). 4.1.4 Digital people: Growth in a World of Ems

#### AGI is key to stop nanotech and bioterror risks

Turchin and Denkenberger 20 [Alexey Turchin, contributing author on IEET and graduate from Moscow State University where he studied Physics and Art History, and David Denkenberger, associate professor at the University of Canterbury in mechanical engineering with a Ph.D. in Civil Engineering from University of Colorado who studies existential risk, 2020, “Classification of global catastrophic risks connected with artificial Intelligence,” Springer, https://link.springer.com/article/10.1007/s00146-018-0845-5]/Kankee

3.6 Opportunity cost of not preventing other existential risks Other global risks could appear if superintelligent AI does not emerge in time to prevent them (Bostrom 2003a). Superintelligent AI and its supposed ability to control many parameters and predict the future is our best chance of avoiding the risks of mature biotechnology and nanotechnology (Yudkowsky 2008). Without superintelligent AI, humanity may not be able to control the dissemination of dangerous biotechnologies, which will be available to thousands of potential biohackers, who could create thousands of pathogens and produce a global multipandemic (Turchin et al. 2017). Thus, if the creation of a powerful and global control system is delayed for decades, perhaps because of a fear of superintelligence, it will increase other GCRs. A global control system would most likely require some form of limited superintelligence, like the AI Nanny suggested by Goertzel (2012). 3.7 AI gains strategic decisive advantage without self‑improving

#### AGI improves hydro-modeling – that’s key to water conservation and climate adaptation

Zhan et al. 25 [Chuanjun Zhan, researcher at the School of Environmental and Municipal Engineering at the Qingdao University of Technology, Zhenxue Dai, researcher at the School of Environmental and Municipal Engineering at the Qingdao University of Technology, Jiu Jimmy Jiao, researcher at the Department of Earth Sciences The University of Hong Kong, Mohamad Reza Soltanian, researcher at the Departments of Geosciences and Environmental Engineering at the University of Cincinnati, Huichao Yin, researcher at the Department of Plant & Environmental Science at New Mexico State University, and Kenneth C. Carroll, researcher at the Department of Plant & Environmental Science at New Mexico State University, 01-18-2025, “Toward Artificial General Intelligence in Hydrogeological Modeling With an Integrated Latent Diffusion Framework,” Geophysical Research Letters, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL114298]/Kankee

Plain Language Summary Hydrogeological modeling is critical for water resource management and environmental protection, but their complex and variable nature makes accurate modeling challenging. Traditional approaches rely on separate models for tasks such as simulating water flow, predicting solute transport, or identifying subsurface structures, and each task requires significant time and effort for setup and adjustment. This study introduces an Latent Diffusion Model (LDM) framework that efficiently performs multiple hydrogeological modeling tasks. The LDM generates detailed representations of subsurface heterogeneity, efficiently predicts flow and mass transport, and directly identifies underground structures from observed data without iterative simulations. By integrating these capabilities into a single framework, the LDM streamlines workflows, reduces computational demands, and improves adaptability across diverse scenarios. This innovative approach not only enhances modeling efficiency but also lays the foundation for intelligent systems to address complex environmental challenges. 1 Introduction Artificial General Intelligence (AGI) refers to a system that can understand, learn, and apply knowledge across a wide range of tasks, exhibiting cognitive abilities similar to those of humans. While significant progress has been made in narrow AI, true AGI remains an elusive goal. Key characteristics include the ability for autonomous reasoning, transfer learning, and domain-independent problem-solving (Fei et al., 2022; Kuusi & Heinonen, 2022). Among various capabilities required for AGI development, multi-task processing represents an essential component, particularly in domains like hydrogeological modeling where integrated approaches are needed to address complex, interconnected problems. The development of AGI in hydrogeological modeling could revolutionize how we approach subsurface characterization, enabling systems that can seamlessly transition between different modeling tasks, adapt to new scenarios without extensive retraining, and integrate diverse data sources in ways that mimic human expertise. Such capabilities would be particularly valuable for addressing challenges like real-time aquifer management, complex contamination scenarios, and climate change impacts on groundwater systems. This transformative vision holds significant potential for numerous fields, including environmental and earth sciences, where the complexity and variability of systems (e.g., groundwater flow and subsurface contaminant migration) necessitate more integrated and flexible modeling approaches. Accurate modeling of groundwater flow, solute transport, and aquifer heterogeneity is essential for effective water resource management and environmental protection (Harp et al., 2008; Kitanidis, 2015; Rajaram & Gelhar, 1995; Rizzo & de Barros, 2019). However, the inherent complexity of subsurface environments, characterized by multiscale heterogeneity and limited observational data, presents significant challenges for traditional modeling approaches (Carrera et al., 2005; Jankovic et al., 2017; Scheibe et al., 2015; Song et al., 2019; Zhu & Yeh, 2005). Existing methods often rely on a series of task-specific models, each optimized for a single objective, such as predicting contaminant transport or describing aquifer structures. While these models have demonstrated utility, they are time-consuming, computationally expensive, and require extensive calibration and optimization for each task (Sbai, 2020; Zhan et al., 2023). Moreover, the need to integrate diverse data sources and types, such as borehole measurements, hydraulic head distributions, and solute concentrations, further complicates traditional modeling workflows. Recent advances in deep learning (DL) have shown promise in addressing some of these challenges by offering more flexible, data-driven approaches to subsurface modeling (Ershadnia et al., 2024; Kang et al., 2021; Mo et al., 2020; J. Zhang et al., 2024). However, achieving different objectives often require distinct model architectures, resulting in a proliferation of task-specific designs. Previous studies frequently employed separate DL models tailored to individual tasks, such as flow prediction or aquifer structure characterization (Cui et al., 2024; Moeini et al., 2024; Zhan, Dai, Soltanian, & Zhang, 2022), without a unified approach. The selection and tuning of models is a time-intensive process, which may offset the efficiency gains promised by DL methods. From a lifecycle perspective, the reliance on multiple specialized models may limit the advantages of DL approaches over traditional computational methods, thereby constraining their broader adoption in hydrogeology to overcome these limitations, the multi-task processing capabilities of AGI present a compelling alternative. AGI enables a single model to handle diverse objectives without requiring task-specific architectures or extensive parameter adjustments (Feng et al., 2024). This integrated approach enhances the overall efficiency of subsurface modeling and represents a significant step toward realizing the vision of AGI in hydrogeological applications. Latent Diffusion Model (LDM), as a next-generation generative artificial intelligence model, offer a promising solution to this challenge. LDM have been widely applied in various image generation tasks (Rombach et al., 2022). Popular models such as Stable Diffusion and DALL·E 3, which power advanced tools such as ChatGPT, are either based on or incorporate LDM (Bengesi et al., 2024). These models excel at a variety of image processing tasks, including generating images conditioned on text or other inputs, modifying specific regions, filling missing areas, and repairing damaged or aged images (Yang et al., 2023). Compared to earlier generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), LDM offer more stable training, enhanced control over conditional generation, and the versatility to handle multiple tasks (Lee et al., 2024). Moreover, unlike other diffusion models such as Denoising Diffusion Probabilistic Models (DDPMs), LDM operate in latent space, significantly reducing training time and computational resource requirements (Blattmann et al., 2023). Inspired by the success of LDM in image generation, their robust training processes, strong conditional control, and multi-task learning capabilities make them highly suitable for addressing various challenges in hydrogeological modeling (Di Federico & Durlofsky, 2024; X. Zhang, Jiang, Wei, et al., 2024). This study investigates the potential of LDM in advancing multi-task processing capabilities for hydrogeological modeling applications. By leveraging the multi-task capabilities of LDM, we aim to address four critical tasks in hydrogeological modeling: generating single-scale aquifer heterogeneity, generating multiscale heterogeneity, constructing surrogate models for groundwater flow and solute transport, and directly identifying aquifer heterogeneity structure from observational data. These tasks are performed within an integrated LDM framework, leading to the development of a versatile hydrogeological modeling framework that facilitates multiscale aquifer characterization and eliminates the need for multiple specialized models. This single-model approach provides a powerful alternative to traditional hydrogeological modeling, enhancing computational efficiency and adaptability across diverse scenarios. Furthermore, this contributes to exploring the multi-task processing capabilities required for future AGI models in hydrogeology. 2 Problem Statement

#### AGI solves every problem and causes billions of years of flourishing

Tegmark 23 [Max Tegmark, a professor of physics and AI researcher at the Massachusetts Institute of Technology, 7-6-2023, "Five ways AI could improve the world: ‘We can cure all diseases, stabilise our climate, halt poverty’", Guardian, https://www.theguardian.com/technology/2023/jul/06/ai-artificial-intelligence-world-diseases-climate-scenarios-experts]/Kankee

5 ‘We can flourish, not just for the next election cycle, but for billions of years’ The positive, optimistic scenario is that we responsibly develop superintelligence in a way that allows us to control it and benefit from it. The “control” part is, I think, more hopeful than many people assume. There is a field of computer science called formal verification, where you come up with a rigorous mathematical proof that a program is always going to do what it’s supposed to. You can even create what is called “proof-carrying code”; it works in the opposite way to a virus checker. If a virus checker can prove that the code you are going to run is malicious, it won’t run it; with proof-carrying code, only if the code can prove that it’s going to do what you want it to do will your hardware run it. This is the type of mechanism we need to ensure advanced AI is safe. We can’t do this yet with GPT-4 or other powerful AI systems, because those systems are not written in a human programming language; they are a giant artificial neural network, and we have almost no clue how they work. But there is a very active research field called mechanistic interpretability. The goal is to take these black-box neural networks and figure out how they work. If this field makes so much progress that we can use AI itself to extract out the knowledge from other AI and see what it has learned, we could then reimplement it in some other kind of computational architecture – some sort of proof-carrying code – that you can trust. Then you can still use the power of neural networks to discover and learn, but now you can trust something that’s way smarter than you. Then what are we going to do with it? Well, the sky’s the limit. We can cure all diseases, stabilise our climate, eliminate poverty, etc. We can flourish not just for the next election cycle, but for billions of years. We have been on this planet for more than 100,000 years, and most of the time we have been like a leaf blowing around in the wind, without much control of our destiny, just trying to not starve or get eaten. Science and technology and human intelligence have made us the captains of our own ship. I find that inspiring. If we can build and control superintelligence, we can quickly go from being limited by our own stupidity to being limited by the laws of physics. It could be the greatest empowerment moment in human history.

#### AGI solves climate change and disease – innovation feedback loops

Kurzweil 24 [Ray Kurzweil, Computer scientist with over 20 honorary doctorates in science and engineering, a BS in computer science from MIT, and 60+ years of experience in AI/computer science research, 06-18-2024, "Ray Kurzweil on how AI will transform the physical world", Economist, https://www.economist.com/by-invitation/2024/06/17/ray-kurzweil-on-how-ai-will-transform-the-physical-world]/Kankee

BY THE TIME children born today are in kindergarten, artificial intelligence (AI) will probably have surpassed humans at all cognitive tasks, from science to creativity. When I first predicted in 1999 that we would have such artificial general intelligence (AGI) by 2029, most experts thought I’d switched to writing fiction. But since the spectacular breakthroughs of the past few years, many experts think we will have AGI even sooner—so I’ve technically gone from being an optimist to a pessimist, without changing my prediction at all. After working in the field for 61 years—longer than anyone else alive—I am gratified to see AI at the heart of global conversation. Yet most commentary misses how large language models like ChatGPT and Gemini fit into an even larger story. AI is about to make the leap from revolutionising just the digital world to transforming the physical world as well. This will bring countless benefits, but three areas have especially profound implications: energy, manufacturing and medicine. Sources of energy are among civilisation’s most fundamental resources. For two centuries the world has needed dirty, non-renewable fossil fuels. Yet harvesting just 0.01% of the sunlight the Earth receives would cover all human energy consumption. Since 1975, solar cells have become 99.7% cheaper per watt of capacity, allowing worldwide capacity to increase by around 2m times. So why doesn’t solar energy dominate yet? The problem is two-fold. First, photovoltaic materials remain too expensive and inefficient to replace coal and gas completely. Second, because solar generation varies on both diurnal (day/night) and annual (summer/winter) scales, huge amounts of energy need to be stored until needed—and today’s battery technology isn’t quite cost-effective enough. The laws of physics suggest that massive improvements are possible, but the range of chemical possibilities to explore is so enormous that scientists have made achingly slow progress. By contrast, AI can rapidly sift through billions of chemistries in simulation, and is already driving innovations in both photovoltaics and batteries. This is poised to accelerate dramatically. In all of history until November 2023, humans had discovered about 20,000 stable inorganic compounds for use across all technologies. Then, Google’s GNoME AI discovered far more, increasing that figure overnight to 421,000. Yet this barely scratches the surface of materials-science applications. Once vastly smarter AGI finds fully optimal materials, photovoltaic megaprojects will become viable and solar energy can be so abundant as to be almost free. Energy abundance enables another revolution: in manufacturing. The costs of almost all goods—from food and clothing to electronics and cars—come largely from a few common factors such as energy, labour (including cognitive labour like R&D and design) and raw materials. AI is on course to vastly lower all these costs. After cheap, abundant solar energy, the next component is human labour, which is often backbreaking and dangerous. AI is making big strides in robotics that can greatly reduce labour costs. Robotics will also reduce raw-material extraction costs, and AI is finding ways to replace expensive rare-earth elements with common ones like zirconium, silicon and carbon-based graphene. Together, this means that most kinds of goods will become amazingly cheap and abundant. These advanced manufacturing capabilities will allow the price-performance of computing to maintain the exponential trajectory of the past century—a 75-quadrillion-fold improvement since 1939. This is due to a feedback loop: today’s cutting-edge AI chips are used to optimise designs for next-generation chips. In terms of calculations per second per constant dollar, the best hardware available last November could do 48bn. Nvidia’s new B200 GPUs exceed 500bn. As we build the titanic computing power needed to simulate biology, we’ll unlock the third physical revolution from AI: medicine. Despite 200 years of dramatic progress, our understanding of the human body is still built on messy approximations that are usually mostly right for most patients, but probably aren’t totally right for you. Tens of thousands of Americans a year die from reactions to drugs that studies said should help them. Yet AI is starting to turn medicine into an exact science. Instead of painstaking trial-and-error in an experimental lab, molecular biosimulation—precise computer modelling that aids the study of the human body and how drugs work—can quickly assess billions of options to find the most promising medicines. Last summer the first drug designed end-to-end by AI entered phase-2 trials for treating idiopathic pulmonary fibrosis, a lung disease. Dozens of other AI-designed drugs are now entering trials. Both the drug-discovery and trial pipelines will be supercharged as simulations incorporate the immensely richer data that AI makes possible. In all of history until 2022, science had determined the shapes of around 190,000 proteins. That year DeepMind’s AlphaFold 2 discovered over 200m, which have been released free of charge to researchers to help develop new treatments. Much more laboratory research is needed to populate larger simulations accurately, but the roadmap is clear. Next, AI will simulate protein complexes, then organelles, cells, tissues, organs and—eventually—the whole body. This will ultimately replace today’s clinical trials, which are expensive, risky, slow and statistically underpowered. Even in a phase-3 trial, there’s probably not one single subject who matches you on every relevant factor of genetics, lifestyle, comorbidities, drug interactions and disease variation. Digital trials will let us tailor medicines to each individual patient. The potential is breathtaking: to cure not just diseases like cancer and Alzheimer’s, but the harmful effects of ageing itself. Today, scientific progress gives the average American or Briton an extra six to seven weeks of life expectancy each year. When AGI gives us full mastery over cellular biology, these gains will sharply accelerate. Once annual increases in life expectancy reach 12 months, we’ll achieve “longevity escape velocity”. For people diligent about healthy habits and using new therapies, I believe this will happen between 2029 and 2035—at which point ageing will not increase their annual chance of dying. And thanks to exponential price-performance improvement in computing, AI-driven therapies that are expensive at first will quickly become widely available. This is AI’s most transformative promise: longer, healthier lives unbounded by the scarcity and frailty that have limited humanity since its beginnings.

#### AGI helps solve neurodegenerative disorders – that saves millions of lives and increases their quality of life

Qadri et al. 24 [Yazdan Ahmad Qadri, researcher at the School of Computer Science and Engineering at Yeungnam University, Khurshid Ahmad, researcher at the Department of Health Informatics, College of Applied Medical Sciences, Qassim University, and Sung Won Kim, researcher at the School of Computer Science and Engineering at Yeungnam University, 10-15-2024, "Artificial General Intelligence for the Detection of Neurodegenerative Disorders", MDPI, https://www.mdpi.com/1424-8220/24/20/6658]/Kankee

1. Introduction Organization for Economic Co-operation and Development (OECD) nations have seen their populations decline along with an increase in average life expectancy [1]. This high average life expectancy, paired with a low fertility rate, results in a rapidly aging population, which threatens the national economy and healthcare systems. Statistics reveal that South Korea is facing the fastest decline in population growth and is on the path to becoming a super-aged society by 2025. With this increase in age, the presumed percentage of the population with dementia is expected to increase, as per a report by the National Health Insurance Service of the Republic of Korea [2]. The incidence of Parkinson’s disease (PD) and Alzheimer’s disease (AD) in South Korea has notably risen, which coincides with existing studies which correlated age with these neurodegenerative disorders (NDs) [3,4]. An estimated 8.5 million people were suffering from PD in 2019 according to the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD), which furthermore caused a 100% increase in the number of deaths between 2000 and 2019 [5]. An estimated 50 million people suffer from AD, while dementia is the fifth leading cause of death across the globe [6]. Females are more likely to suffer from these diseases compared with males [7]. PD is a progressive ND which significantly degrades the quality of life (QoL) due to its impact on motor and locomotive functions. AD is a severe form of dementia which affects behavior, memory, and cognition [3,8]. In OECD countries, the economic burden of these diseases is projected to rise significantly as the population ages [9]. Early and accurate diagnosis of PD and AD is critical in managing symptoms and maintaining a stable QoL. The diagnostic tools for detecting PD and AD include clinical evaluations, imaging, and biomarker analysis [10,11]. The patients undergo an extensive clinical evaluation to identify changes in locomotion and cognitive abilities. On the other hand, radiological tools include magnetic resonance imaging (MRI) and positron emission tomography (PET) scans [12,13]. Genetic testing and biomarker analysis can further consolidate the reliability of the diagnosis of these diseases. Artificial intelligence (AI) can assist in and improve the accuracy of the diagnostic process. The United States Food and Drug Administration (FDA) has approved several AI and machine learning (ML)-based radiological image analysis tools [14], therefore, AI and ML are coming to the forefront in the early detection and diagnosis of NDs. Neural networks (NNs) can accurately detect patterns in biomedical signals and identify objects in medical images. Convolutional neural networks (CNNs) can successfully identify abnormalities in the medial images [15]. Recurrent neural networks (RNNs) can analyze time series data to identify anomalies in the normal functions of physiological functions. Deep learning (DL) algorithms have demonstrated high accuracy in anomaly detection [16]. Additionally, large language models (LLMs) can analyze massive tranches of image and clinical datasets, clinical studies, and the research literature and assist in accurately diagnosing NDs. AI algorithms can analyze medical records, including imaging, genetics, and biomarkers, to identify patterns indicative of PD and AD [17,18]. The hope of artificial general intelligence (AGI) leading to superhuman capabilities in machines is contended in [19]. AGI is the capability of an AI model to mimic human cognitive processes. AGI allows machines to “think” and “learn”, which allows them to understand concepts and apply them across several domains. Therefore, machines can transfer the knowledge they have learned in one domain to another domain. AGI can possess four characteristics. Firstly, they can perform an unlimited number and types of tasks. Secondly, they can generate new tasks within a context, and thirdly, the agents operate using a value system which underpins task generation. Finally, they can visualize the world in a global model, which they can use to interact with the physical world [20]. LLMs are exemplified by models like ChatGPT, Gemini, and Llama, which exhibit the traits of AGI; they can gather knowledge from vast online resources and transfer their learning from these large resources to perform an unlimited number of tasks [21]. These models can perform tasks across domains such as mathematics, language processing, image, video, text generation, medicine, and software coding. Their application in the field of healthcare is profound, ranging from drug discovery to medical image processing, genomics, and clinical assistance [22]. Detecting NDs using next-generation AI models has gained traction in the past few years and attracted significant research interest. The advent of transformer models has enabled AI-based systems to understand the context of data [23]. This ability to contextualize information is critical in establishing correlation between the symptoms and history and arriving at an accurate diagnosis [24]. This review is aimed at providing insight into the prediction, early detection, and diagnosis of NDs using the latest AGI models, with a focus on PD and AD. Radiological data analysis in the context of clinical evaluations and medical tests can assist in triangulating the underlying causes and hence determining a prudent treatment plan. The American College of Radiology Data Science Institute’s AI Central database contains a list of 200 FDA-approved AI products to assist in imaging-based diagnostic products from across 100 manufacturers [25]. Approval from regulators accelerates the integration of computer-aided diagnosis in healthcare systems. A survey of DL-based approaches for detecting NDs was presented in [26]. The survey covered disease detection along with severity analysis, presenting a CNN-based methodology for ND detection. The accuracy of various DL-based approaches in identifying these diseases was between 89% and 97%. The severity of the disease was estimated with a success rate greater than 90%. This work identifies the state-of-the-art AGI methodologies which can diagnose PD and AD using massive repositories of clinical and experimental data, including the research literature. Combining Internet of Things (IoT) with AGI can alleviate and streamline the process of monitoring at-risk individuals for diagnosis of NDs. A primer on PD and AD is presented, describing the diagnostic tools used to predict and diagnose these disorders, in Section 2. The role of AGI in the diagnosis of PD and AD is identified. To the best of our knowledge, this review is the first to present a detailed discussion on the role of AGI in ND diagnosis (Section 3). An IoT-based framework based on our previous work is presented for the ubiquitous monitoring and diagnosis of NDs in Section 4. Section 5 presents a roadmap for the future to identify the avenues for mitigation of challenges in this area. Section 6 concludes the discussion. 2. Background 2.1. Neurodegenerative Disorders

#### AGI solves every crisis – medicine, climate, disease, poverty, economic growth, inequality, and war

Losey 24 [Ralph Losey, Adjunct Professor of Law at the University of Florida with a JD from the University of Florida and a BA from Vanderbilt in philosophy, 08-08-2024, “Artificial General Intelligence, If Attained, Will Be the Greatest Invention of All Time”, EDRM, https://www.jdsupra.com/legalnews/artificial-general-intelligence-if-8762682/]/Kankee

A computer capable of Artificial General Intelligence (AGI), if ever created, will be the greatest invention in history. It will be far more transformative than the printing press or the wheel. Unlike narrow AI, which excels in specific tasks, AGI will have a general ability to understand, learn, and apply knowledge at a level equal to or beyond human intelligence. If AGI is attained in a computer, it will be the greatest invention of all time, primarily because of its impact on all facets of human life and civilization, including our systems of justice. AGI Defined Artificial General Intelligence means a computer system with a level of AI that is equal or greater than human intelligence in all fields. Narrow AI, which has intelligence to perform specific limited tasks, has already attained greater than human intelligence in some fields. The fields include games, such as Chess and Go, facial recognition, translation, and scientific data-heavy areas, such as protein folding. AGI, when invented, will be capable of performing at a human or better level in all fields of knowledge and all intellectual tasks, including math, science, coding, general reasoning, problem-solving, planning, and adapting to new situations. You name it, it will be as smart or smarter than the top human minds in that field. Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Microsoft Research, 4/13/23) (“Given the breadth and depth of GPT-4’s capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system.”) This broad cognitive capability will enable AGI to integrate knowledge from various domains, make autonomous decisions, and continuously improve its performance through learning and experience. If attained, AGI computers will revolutionize human knowledge and fundamentally transform all human society and culture, including law. Ray Kurzweil and others believe AGI will be attained in five years, 2029. Ray Kurzweil: Google’s prophet of superintelligent AI who will not slow down (e-Discovery Team, 12/12/23). Some think it may come sooner. Some think that AI self-awareness will result. I do not, but I am confident AGI will have advanced metacognition abilities. Bill Gates on the Next ‘Big Frontier’ of Generative AI: Programming Metacognition Strategies into ChatGPT (e-Discovery Team, 7/26/24). Ray Kurzweil predicts that once AGI is invented by a company such as Google or OpenAI, it will quickly spread and be inexpensively available to everyone, much like generative AI is now. He also predicts this will then lead to the Singularity in 2045, where humans will merge with superintelligent AI, leading to a million-fold increase in our intelligence. Ray Kurzweil’s New Book: The Singularity is Nearer (when we merge with AI) (e-Discovery Team 7/17/24). Others believe that AGI will never be attained or is hundreds of years away. Based on Ray Kurzweil’s excellent track record, strong arguments, and position at Google, it seems reasonable to take him seriously and start preparing for AGI now. Id. Unmatched Problem-Solving Capabilities AGI, when attained, should be able to address complex global challenges that have persisted for centuries. Its advanced cognitive abilities should enable it to: Analyze Vast Data Sets: AGI can process and analyze massive amounts of data far more quickly and accurately than humans, leading to breakthroughs in fields such as medicine, climate science, and economics. Generate Innovative Solutions: By integrating knowledge from various disciplines, AGI can develop innovative solutions to problems like climate change, pandemics, and resource scarcity. Optimize Systems: AGI can optimize systems ranging from healthcare to transportation, improving efficiency and reducing waste on a global scale. Exponential Advancement in Knowledge and Technology AGI is expected to catalyze exponential growth in scientific and technological progress: Accelerated Research: AGI can conduct research at a pace unimaginable for human scientists, rapidly advancing our understanding of the universe, biology, and more. Innovation in Technology: AGI-driven innovation will lead to the creation of new technologies and the enhancement of existing ones, driving economic growth and improving quality of life. Interdisciplinary Integration: AGI can seamlessly integrate knowledge from diverse fields, leading to holistic advancements that human researchers might miss due to the limitations of siloed disciplines. The legal industry will also benefit immensely from AGI-driven advancements. Automated contract analysis, predictive policing, and enhanced legal analytics are just a few areas where AGI could significantly enhance efficiency and effectiveness. As I often note on the e-Discovery Team blog, the integration of advanced AI technologies in legal processes is not just beneficial but necessary for the legal profession to keep pace with the rapidly evolving technological landscape​​. Enhanced Human Capabilities and Quality of Life AGI has the potential to significantly enhance human capabilities and overall quality of life: Personalized Education: AGI can provide individualized learning experiences, catering to each person’s strengths and weaknesses and ensuring everyone reaches their full potential. In the legal sector, AGI could enhance legal education by providing personalized learning pathways for law students and continuous professional development for practicing lawyers. AGI could tailor legal training programs to address individual learning gaps, ensuring a more competent and knowledgeable legal workforce. Advanced Healthcare: AGI can revolutionize healthcare by diagnosing diseases with unprecedented accuracy, personalizing treatments, low-cost medicines, and even predicting health issues before they arise. Economic Transformation: AGI can drive economic growth by automating mundane tasks, allowing humans to focus on creative and strategic activities, creating low-cost goods, foods, energy, housing and significantly reducing poverty and inequality. Ethical and Societal Impacts The ethical implications of AGI are significant, and if managed responsibly, AGI can lead to a more just and equitable society: Fair Distribution of Resources: AGI can optimize the distribution of resources, ensuring fair access to essentials like food, water, and healthcare. Global Collaboration: AGI can facilitate international cooperation by providing unbiased analysis and recommendations, helping to resolve conflicts and promote peace. Ethical Frameworks: With proper governance, AGI can be programmed to uphold ethical standards, ensuring that its actions are aligned with human values and societal good. Historical Context and Long-term Impact Comparing AGI to previous landmark inventions highlights its unique potential: The Wheel: Revolutionized transportation and industry. The Printing Press: Democratized knowledge and spurred the Scientific Revolution. Many believe it is the most important invention of all time, at least before AGI. Electricity: Powered modern civilization and technological advancements. The Internet: Connected the world and transformed communication and information access. AGI will encompass the benefits of these inventions and surpass them by offering not just tools but an intelligent entity capable of continuous learning and improvement. That will lead to sustained progress in all fields of knowledge and endeavors. Profound Transformations of the Legal Profession from AGI The legal profession, once thought to be slow to adopt new technologies, stands to be profoundly transformed by AGI. The potential applications are vast and varied. Many in the legal profession are already catching on and using the narrow AI we have today. The stodgy old times of a slow-moving legal profession are already dissolving, with e-discovery specialists leading the way. Predictive Analytics AGI can provide very powerful predictive analytics to forecast case outcomes. This helps lawyers to strategize more effectively. By analyzing historical data, case law, and judge-specific decision patterns, AGI can predict the likelihood of various outcomes in litigation. Our existing level AI is already very proficient at this. With even more powerful AGI, negotiation strategies will improve and lead to ever more judicious uses of litigation resources. Document Review and E-Discovery Document review and e-discovery are critical, labor-intensive aspects of modern legal practice. AI technologies like predictive coding have already begun to revolutionize these areas. AGI will take this to the next level, automating these processes with even greater accuracy and speed. Traditional document review involves sifting through vast quantities of documents to identify those relevant to a case, a task that is time-consuming and prone to human error. AGI can automate this process even further, rapidly identifying relevant documents, categorizing them by importance, and flagging potentially privileged information. Moreover, AGI can learn from each case, improving its accuracy and efficiency over time. This automation not only reduces costs but also ensures a more thorough and reliable review process, allowing legal professionals to focus on higher-value tasks. Legal Research Legal research is another area poised for further transformation by AGI. Currently, lawyers spend considerable time researching case law, statutes, and legal precedents to support their arguments. AGI can perform comprehensive legal research in seconds, providing lawyers with relevant case law, statutes, and legal precedents, thereby saving time and reducing costs. AGI’s ability to cross-reference vast legal databases instantaneously means that no relevant case or statute is overlooked. Additionally, AGI can present the findings in a concise and understandable manner, complete with contextual analysis and relevance ranking. This allows lawyers to quickly access the most pertinent information, enhancing the quality and precision of their legal arguments.

#### AGI is key to improve health outcomes with better, customized, and earlier treatments and diagnoses – heart disease proves

Haq et al. 25 [Rashid Ul Haq, researcher at the Abdul Wali Khan University Mardan, Hashim Ali, researcher at the Abdul Wali Khan University Mardan, Mehak Mushtaq Malik, researcher at COMSATS University Islamabad, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Mariya Ouaissa, researcher at the Computer Systems Engineering Laboratory at Cadi Ayyad University, Mariyam Ouaissa, researcher at the Laboratory of Information Technologies at Chouaib Doukkali University, and Inam Ullah Khan, former visiting researcher at King’s College London and General Chair at the International Conference on Trends and Innovations in Smart Technologies with a Ph.D. in Electronics Engineering from Isra University and a Bachelor of Computer Science from Abdul Wali Khan University, 2025, “Chapter 8 Review of Heart Disease Prediction Using AGI Models: Advancements and Challenges Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Future Directions and Research Opportunities There are various issues and restrictions in current research on heart disease prediction. As the models are still not enough to be deployed to the general population, further research is needed to improve their accuracy and reliability. Future research on heart disease prediction might benefit from the use of Artificial General Intelligence (AGI) since it could increase accuracy and dependability. The analytical powers of AGI can find new risk variables and improve model performance. This integration provides the potential to discover important insights for more accurate prediction and opens up new areas for investigation. AGI has the potential to anticipate and prevent cardiac disease with greater accuracy and dependability. Integration of Multimodal Data for Comprehensive Prediction As a consequence of technological improvements, a number of hospitals and clinics have recently acquired patient data; however, they do not make it available on the internet or the quality is inadequate, so the researcher continues to use outdated data. As a result, increasing the availability and quality of patient data may allow us to develop more effective algorithms and gain a better understanding of the risk factors for heart disease. Modern technologies and tools are decreasing the error rate and hence will be helpful in the future with better prediction [20]. Data rate, power constraints, and many other factors like these are delicate to deal with and are required to be handled carefully [21]. Similarly, if the input consists of images and heart sounds, the model can be trained to recognize trends and alert users to potential concerns before they become serious [22]. This could lead to earlier detection and treatment of cardiac illness, as well as better patient outcomes and cheaper healthcare expenditures. There is promise for more accurate and early identification, better patient outcomes, and efficiency when Artificial General Intelligence (AGI) is integrated with multiple modalities for complete prognosis in cardiac disease. The capacity of AGI to examine many kinds of data, including pictures and heart sounds, can improve early warning systems and trend identification. Along with addressing issues with data availability and quality, this integration improves prediction models and helps identify risk indicators. Healthcare costs can be lowered by using AGI through preemptive interventions and efficient resource allocation. AGI holds the potential to provide more accurate and effective comprehensive heart disease prediction. Personalized Medicine Approaches in Heart Disease Prediction As the algorithms advance, they may be able to add individual patient information like age, gender, and lifestyle aspects to produce even more personalized forecasts. These models could also be coupled to wearable devices or mobile apps for easy and continuous monitoring of heart health. With more accurate and customized predictions, the use of Artificial General Intelligence (AGI) in personalized medicine techniques for cardiac disease prediction has great potential to improve patient care. We can improve the precision and efficacy of forecasts, enabling proactive therapies and enabling people to actively control their heart health, by combining specific patient data and utilizing AGI’s analytical skills. Improved patient outcomes might result from this combination, which has the potential to transform personalized therapy in heart disease prediction. Conclusion The goal of this review study is to provide helpful insights into the advancements, limitations, and potential of heart disease prediction algorithms. It aims to contribute to ongoing research and development efforts in heart disease prediction by critically analyzing the effectiveness and limitations of existing AGI models, resulting in improved early detection, prevention, and management of this vital health concern. The purpose of this study is to discover limitations and potentials in the categorization of heart disease literature. The accuracy of the algorithms was used as a comparison parameter. The number of data points utilized in the model, the number of attributes used in the model, and the preprocessing processes used in the specific study were discovered to have the largest impact on accuracy. It also depends on the algorithms used, which are determined by the dataset’s attributes. In the majority of research studies, the ANN technique has the highest accuracy on the Cleveland dataset. With the exception of ANN, SVM and random forest outperformed the other approaches. Our research and study show that SVM performs best when datasets can be divided linearly; however, alternative techniques, like MLP, may perform better when datasets cannot be separated linearly. Before deciding on the optimal machine learning technique, the attributes of the dataset must be evaluated. As a result, before choosing the best AGI approach, it is necessary to thoroughly examine the dataset and its properties. The predictive power of heart disease algorithms can be greatly enhanced by AGI’s capacity to decipher intricate patterns, manage missing values and outliers, and utilize a larger variety of data sources and attributes. The constraints of linear separation can be addressed and complex relationships within the dataset can be captured by combining AGI with conventional statistical models. Furthermore, by taking into account unique patient features and customizing therapies accordingly, AGI can support personalized medicine. But it’s imperative to integrate AGI sensibly, taking ethical issues like algorithmic bias, data security, and privacy into consideration. The incorporation of AGI has the potential to completely transform the prediction of cardiac disease, resulting in better early detection, more precise diagnosis, and better management of this serious health issue with careful thought and appropriate application.

#### Humans are comparatively less accurate and speedy, resulting is delays, misdiagnoses, and death

Haq et al. 25 [Rashid Ul Haq, researcher at the Abdul Wali Khan University Mardan, Hashim Ali, researcher at the Abdul Wali Khan University Mardan, Mehak Mushtaq Malik, researcher at COMSATS University Islamabad, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Mariya Ouaissa, researcher at the Computer Systems Engineering Laboratory at Cadi Ayyad University, Mariyam Ouaissa, researcher at the Laboratory of Information Technologies at Chouaib Doukkali University, and Inam Ullah Khan, former visiting researcher at King’s College London and General Chair at the International Conference on Trends and Innovations in Smart Technologies with a Ph.D. in Electronics Engineering from Isra University and a Bachelor of Computer Science from Abdul Wali Khan University, 2025, “Chapter 8 Review of Heart Disease Prediction Using AGI Models: Advancements and Challenges Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Introduction The heart is considered an important organ in humans. It moves blood around the body, supplying all organs and tissues with nutrients and oxygen. Without a heartbeat, the body cannot survive, and insufficient blood flow has the potential to be fatal [1]. As a result, if it is found early on, there is less chance of death and more of a chance that the treatment will be effective. Early detection can lower the mortality rate. Traditionally, a doctor would examine a patient’s symptoms and medical history before deciding which tests to order. However, this conventional method can be time-consuming and may cause delays or missed chances for early intervention Additionally, it is possible to overlook specific symptoms or risk factors that could result in an incorrect diagnosis. To address these research concerns, there is growing interest in the application of AI and machine learning algorithms to aid in medical diagnostics. These technologies can rapidly analyze massive volumes of data and find trends that human clinicians may not see right away, resulting in more accurate and timely diagnoses. Medical diagnostics may be revolutionized by integrating Artificial General Intelligence (AGI) into heart disease prediction. Machine learning algorithms can rapidly analyze huge amounts of patient data and detect deep relationships and patterns by utilizing artificial general intelligence (AGI) and its capacity to absorb, acquire, and apply knowledge across several domains. This comprehensive exam, which takes into factors other than a patient’s medical history and symptoms, may help make more accurate heart disease forecasts. Furthermore, by taking into account unique patient features and customizing treatment strategies accordingly, AGI may make personalized medicine possible. To guarantee the proper application of AGI in healthcare, ethical issues like algorithmic bias, privacy, and data security must be carefully taken into account. In the field of cardiology, AGI has the potential to greatly increase the precision of diagnosis while enhancing patient results with the right research, development, and cooperation. A thorough analysis of the methods used to predict heart problems is discussed in this review. We will also go over the possible advantages and drawbacks of using these algorithms, as well as the moral questions raised by doing so. Understanding how these technologies can enhance patient outcomes is crucial, as is making sure they are applied responsibly and ethically. The research is mainly focused on Ann’s comparative performance in relation to others and primary dataset properties or characteristics that affect the algorithm selection. This research aims to shed light on the best algorithm for reliably forecasting heart disease. Healthcare professionals can improve patient outcomes by making better decisions and identifying the critical variables that affect algorithm selection. The remaining sections of this review are divided into seven categories: Data Sources and Features using AGI, which includes data source information, Traditional statistical models explaining the conventional models employed for heart prediction, and similar machine learning techniques addressing the sophisticated techniques employed to predict heart disease. After exploring through evaluation metrics and validation techniques, several measures are used to analyze and evaluate the performance of these models. The section on discussion and limitations offers information on potential drawbacks and difficulties in applying these models to the prediction of heart disease. Future directions and opportunities for research in this area are also highlighted. In conclusion, this essay offers a thorough analysis of the various methods for predicting heart disease, highlighting both their strengths and weaknesses. Data Sources and Features Using AGI Some of the data sources used in this study include laboratory test results, patient surveys, and medical records. The characteristics that were looked at included age, gender, family history of heart disease, lifestyle factors like smoking and exercise habits, and various clinical measurements like blood pressure and cholesterol levels. There is great potential for improving our knowledge and prediction skills when Artificial General Intelligence (AGI) is applied to the issue of data sources and characteristics for heart disease prediction. A deeper understanding of the intricate variables impacting heart disease may be attained by utilizing AGI’s processing and interpretation capabilities for a variety of data sets, such as test results from laboratories, patient questionnaires, and medical records. AGI is able to analyze various data sources more thoroughly and effectively, spotting complex links and patterns that human physicians would not see right away. By taking into account a variety of parameters, including age, gender, family history, lifestyle choices, and clinical data like blood pressure and cholesterol levels, this integration makes it possible to take a more comprehensive approach to risk assessment. We can increase the precision and accuracy of cardiac disease prediction by utilizing AGI, which will result in more individualized treatments and better patient outcomes. Source of Data Utilized in Predicting Heart Disease and Relevant Features This classification makes use of the UCI Cleveland database’s data on heart disease. The dataset has 303 tuples and 76 characteristics. Age, ca, cholesterol (mg/dl), chest pain type, exang, fbs (fasting blood sugar), oldpeak, restecg, and trestbps (mmHg) are used in the majority of papers. We can use AGI’s processing and analysis capabilities to handle and evaluate the massive volume of data on heart illness by combining it with data from the UCI Cleveland database. The capacity of AGI to decipher intricate patterns and correlations in the data might yield important insights into the pertinent characteristics for heart disease prediction. By using AGI, we can find previously undiscovered relationships between clinical data such as fasting blood sugar, kind of chest pain, age, cholesterol, and other factors, which can help us create more precise and individualized risk assessments. By enabling more focused and accurate therapies, this combination has the potential to completely transform the prediction of heart disease. This would ultimately enhance the well-being of patients and reduce the impact of heart disease on both individuals and healthcare systems. Data Preprocessing Techniques for Handling Missing Values and Outliers Cleaning data, handling missing values, and modifying data to make them suitable for analysis are all tasks that are included in the process of preparing data for an algorithm, also known as data preparation or preprocessing [2]. In the case of the Cleveland dataset, standardization is required because every feature in the dataset has a different range. It will be possible to ensure that each feature contributes fairly to the output by standardizing the range of all features in the Cleveland database [3]. The dataset contained some missing values, which interpolation values were used to fill in. There is a lot of promise for enhancing the accuracy and dependability of data analysis when Artificial. General Intelligence (AGI) is applied to the problem of data preparation methods for managing outliers and missing values. The detection and management of missing values can be significantly improved by AGI’s capacity to analyze complicated patterns and connections within the data. AGI can be used to automatically fill in values that are missing using sophisticated interpolation algorithms that take into consideration the context and patterns present in the dataset. Additionally, by recognizing abnormal data items that differ noticeably from the norm, AGI can help with outlier detection and management. By guaranteeing that the input data is clear, consistent, and appropriate for precise analysis, this AGI integration might expedite the data preparation stage. We can increase the robustness and dependability of heart disease prediction models by utilizing AGI, which will result in more precise diagnoses and better patient outcomes. Traditional Statistical Models for Heart Disease Prediction Decision trees [4] and logistic regression [5] are two examples of the conventional statistical models that are applied to the Cleveland dataset. The complexity of the heart disease dataset and the limitations of conventional statistical techniques, however, raise the possibility that these models are not always accurate. The heart disease dataset’s complex patterns and linkages may be analyzed using AGI in ways that standard statistical methods cannot. Combining AGI enables us to create more complex and adaptable models that better capture nonlinear linkages and complex interactions, improving our ability to make predictions. To improve the customization of risk assessments, AGI-powered algorithms may make use of a broader range of characteristics and variables, including genetic information, lifestyle variables, and medical imaging. By offering more precise diagnoses and customized therapies, this integration has the potential to completely transform the prediction of cardiac disease and eventually result in better patient outcomes and more efficient healthcare practices.

#### AGI is key to life-saving healthcare innovation and treatment accuracy

Fahad et al. 25 [Muhammad Fahad, researcher at the National University of Computer and Emerging Sciences, Tayyaba Basri, researcher at the National University of Computer and Emerging Sciences, Muhammad Ameer Hamza, researcher at COMSATS University Islamabad, Sheikh Faisal, researcher at the National University of Computer and Emerging Sciences, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Usman Haider, researcher at the National University of Computer and Emerging Sciences, and Salma El Hajjami, Assistant Professor and Researcher at the Faculty of Science at Ibn Zohr University, member of the International Association of Engineers, and member of the International Association of Online Engineering with a Ph.D. in 2021 in Computer from Sidi Mohammed Ben Abdellah University, 2025,“Chapter 2 The Benefits and Risks of Artificial General Intelligence (AGI),” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Artificial General Intelligence Applications: Artificial General Intelligence (AGI) is a significant breakthrough with promising capabilities outperforming traditional AI systems. AGI can potentially cause transformative development in various domains [15]. In the Healthcare Sector, AGI has the potential to be a game changer with the capacity to significantly enhance the diagnostic process by providing quick and precise insights into various medical conditions. AGI can help medical personnel make more accurate and fast diagnoses by evaluating large datasets such as patient records, medical images, and clinical research. Furthermore, AGI can assist scientists in discovering new medicines by pretending how different medications might work. This could speed up the development of critical life-saving medications [38]. The incorporation of Artificial General Intelligence (AGI) into education changes the way we teach and learn. Intelligent tutoring systems that adjust to each student’s pace and preferences can be created using AGI, making learning more engaging. It also improves assessment and feedback by providing detailed insights into student performance. The language understanding capabilities of AGI enable peer-to-peer learning that fosters critical thinking and problem-solving skills. AGI in education, in short, provides unique, flexible, and compelling learning experiences that prepare students for a rapidly evolving world [1]. AGI is beneficial in scientific discovery because it uses intelligent concepts such as Mathematization, Optimization, Analogies, Concept Combination, and Universality. AGI helps scientists by applying these ideas to solve tricky issues and discover new insights in various fields. AGI systems use these concepts to solve complex problems, uncover hidden structures, and generate insights across multiple domains. It’s like having a super-smart assistant who helps us learn more about the world and speeds scientific discovery [39]. Other applications of AGI include its success in gaming, where it has demonstrated its abilities by winning complex games like “Go”. AGI’s prowess is shown by notable examples such as DeepMind’s Alpha Go defeating human champions. In the military, AGI plays a critical role in technologies such as unmanned aerial vehicles and self-governing fighters, which have the potential to revolutionize the way wars are battled. AGI also excels at content creation, producing large quantities of written scripts, art, and video games. It can be used to create conversational AI, such as intelligent personal assistants and chatbots that converse like humans. AGI continues to drive advancements in the automotive industry, allowing for the creation of fully self-driving vehicles. These applications demonstrate AGI’s potential to transform various industries, including entertainment, art, defense, and transportation [40]. These applications demonstrate AGI’s multifaceted capabilities and potential to address complex challenges across multiple domains, indicating both opportunities and ethical concerns in the era of advanced artificial intelligence [10]. Language Model Based on AGI

#### AGI telehealth improves diagnostic treatment and reduces healthcare costs

Asif et al. 25 [Ali Asif, researcher at the National University of Sciences and Technology, Hassan Asif, researcher at the Pakistan Institute of Engineering and Applied Sciences, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Maqsood M. Khan, researcher at the ,Shahzad Latif, researcher at the Department of Computer Science at Szabist University, Muhammad Ameer Hamza, researcher at the Department of Civil Engineering at COMSATS University Islamabad, and Abdur Rehman Khan, researcher at the Air University in Pakistan, 2025, “ Chapter 16 AGI-Enabled Robotics for Healthcare Industry,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Transforming Healthcare Industry Using AGI-Based Robots The healthcare industry has long been marked with complexity and intricate treatment procedures. With the advent of technologies like AGI and ML, this sector is transforming as smart algorithms provide innovative solutions to long-standing challenges. AGI-based robots are assisting medical professionals in many ways. The use of externally wearable sensors provides useful data regarding the health and wellness condition of the user. Such devices are also useful for providing health alerts for the user based on the collected data. They also allow medical professionals to remotely monitor the patient and provide real-time feedback and consultation [60]. This transformation, however, also comes with some challenges. Privacy concerns arise as users’ data is used to gather useful output from the ML algorithm. Other concerns include ethical issues and the risk of system failure. These challenges must be tackled in order for there to be more trust and widespread implementation of the use of such technologies in the healthcare industry. Such transformation has good implications for health care as well. Studies show that AGI technology can diagnose skin cancer more accurately than a professional dermatologist [61]. AGI-based systems are now also being used to diagnose breast cancer [62]. Such examples show how AGI can introduce novel and useful applications in the healthcare industry and make the process of patient monitoring and treatment more efficient. The use of AGI has economic implications for the healthcare industry as well. In most cases, technology-backed solutions prove to be more cost-efficient as compared to traditional methodologies. As an example, Grady Hospital, a public hospital in Atlanta, USA, saved over **$4 million** over two years by implementing an AGI-enabled tool in their workflow [62]. The implementation of AGI-based robotics in health care means that patients can perform routine consultations remotely and get instant help from a medical professional from the comfort of their homes. Expensive visits for routine checkups are also no longer needed, as the diagnosis can be performed from a distance with the help of mobile robots and sensors. These sensors can collect data such as heart rate, blood pressure, and oxygen levels, providing useful insights into the patient’s health condition. AGI-based mobile robots are being used to assist elderly people as well [63], especially those living alone. These robots offer companionship, reminders for medication and daily tasks, and even assist with basic household chores. Their presence helps reduce loneliness and enhances the overall well-being of elderly individuals while providing peace of mind to their families and caregivers. In conclusion, the healthcare industry is undergoing a remarkable transformation with the integration of AGI and ML technologies. AGI-based robots and wearable sensors are revolutionizing patient care, enabling remote monitoring, early intervention, and efficient diagnosis. However, this transformative journey is not without its challenges, including privacy concerns, ethical considerations, and the risk of system failures, which must be addressed to build trust and ensure widespread adoption of these technologies. Nevertheless, the benefits are substantial, with AGI systems demonstrating superior diagnostic accuracy in areas like skin and breast cancer. Moreover, the economic implications are significant, as evidenced by cost savings in institutions like Grady Hospital. AGI-based robotics offer the promise of more accessible, cost-effective, and efficient health care, allowing patients to receive expert care from the comfort of their homes, and providing companionship and support to the elderly. As these technologies continue to evolve, they hold the potential to revolutionize healthcare delivery and improve the overall well-being of individuals around the world.

#### AGI prevents cyber-attacks with better resilience, identification, and threat removal

Arshad et al. 25 [Hira Arshad, Faculty of ICT at the Baluchistan University of Information Technology Engineering and Management Sciences, Ahthasham Sajid, researcher at the Department of Cyber Security and Data Science at Riphah Institute of Systems Engineering at Riphah International University, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Mehak Mushtaq Malik, researcher at the Department of Computer Science COMSATS University Islamabad, and Shahzad Latif, researcher at the Department of Computer Science at the Szabist University Islamabad Campus, 2025, “Chapter 6 A Survey on Cyber Security Encounters and AGI-Based Solutions,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Cyber Ethics Using Concepts of AGI The benefits of AGI can significantly improve cyber security by preventing different types of targeted attacks on the system. Artificial Intelligence (AI) methods, including neural networks and deep learning models, have demonstrated efficacy in recognizing and categorizing data packet attributes, finding irregularities in network traffic, and identifying malevolent actions. Scientists have suggested artificial intelligence (AI)- based intrusion detection systems that make use of machine learning techniques like SVM and ANN to identify botnet attacks and guarantee the security of authentication. AI has also been used to identify anomalous traffic patterns, such as Distributed Denial of Service (DDoS) attacks, by utilizing methods like SVM classification and K-means clustering. It has also been investigated whether AI and parallel computing can work together to speed up model training and increase the speed and effectiveness of network threat detection. All things considered, artificial intelligence (AI) is essential for maintaining overall cyber security, delivering network situational awareness, and promptly identifying high-risk behavior. The advent of Artificial General Intelligence (AGI) presents a significant opportunity to enhance cyber security protocols. AI systems may become more adept at identifying and thwarting cyber-attacks thanks to AGI’s capacity to comprehend and learn from complicated data patterns. Cyber defenses can grow increasingly resilient and aggressive in detecting new attack vectors and adjusting to changing threats by utilizing the cognitive skills of artificial intelligence (AI), such as deductive reasoning, problem-solving, and learning via adaptation. AGI can make more complex anomaly detection algorithms possible, increase threat categorization accuracy, and fortify network defences all around. AGI and AI technology integration can also result in the creation of intelligent and autonomous cyber security systems that are able to constantly evolve, change, and react in real-time to new and emerging cyber threats. AGI has the ability to completely transform cyber security and offer a more proactively and robust defence against increasingly sophisticated cyber-attacks as it develops. Since the Internet is thought of as the largest library in the world and contains knowledge on every subject and field of study, it is usually essential to manage this material with the greatest accuracy and legality. • Avoid logging into other accounts with compromised credentials. • Never give anybody access to your data since there is a potential that they may misuse it and you would find yourself in a lot of trouble. • Never attempt to infect other machines by delivering stealth software to them. Changing permissions, privileged users, and backup adjustments [5, 5] Nature of Cyber Security on Technologies Cyber Security is crucial because it safeguards every kind of information against thieving and loss. Sensitive knowledge protected health info (PHI), in-person specifiable info (PII), property, personal info, data, and government and business info systems square measure all enclosed [7]. Intended Use of Cyber Security According to the Cyber Security & Infrastructure Security Agency, “Cyber security is the art of securing networks, devices, and knowledge from unauthorized access or unlawful usage” (CISA). “The process of guaranteeing data’s confidentiality, integrity, and availability” is another aspect of cyber security [8]. Origin of Cyber Security for Technologies After research assistant Bob Thomas developed the computer virus known as Creeper, which could travel over the ARPANET network, Cyber Security testing was put into place in the 1970s. The pioneer of email, Ray Tomlinson, created the Reaper software, which stalked users and eliminated Creepers [9]. Types of Cyber Security Being Used There are three basic types of cyber security: physical security, cloud security, and network security. The operating systems and specifications you use determine the security of your network. Network protocols, hosts, servers, wireless access points, and firewalls will all be part of it [10]. Data Content of Cyber Security We outline “content” because of the defensive knowledge that our security operations specialists produce to harm a customer’s network against attacks [11]. Ensuring online records are protected from unauthorized exposure, fraud, or theft is known as data security. Ensuring that personal data is protected from fraud, tampering, and unauthorized access is known as data security. This idea encompasses all facets of data security, from the actual safety of gadgets and data storage equipment to the intellectual safety of computer programs, as well as body and access controls. Structured policies and procedures are also included [12]. Primary Cyber Security Threats Elements Infrastructure that Requires AGI-Based Protection Artificial Intelligence (AI) and Artificial General Intelligence (AGI) technologies have shown to be beneficial tools for improving cyber security protocols. These technologies may efficiently identify and categorize data packet characteristics, detect abnormalities in network traffic, and identify malicious behaviors by utilizing neural networks, deep learning models, and artificial general intelligence (AGI) to comprehend complicated data patterns. AI and AGI together allow for the creation of intelligent, self-governing cyber security systems that can constantly learn, adapt, and react in real time to new threats. These systems may dynamically detect and reduce hazards, guaranteeing the safety and stability of electronic structures, by utilizing the reasoning abilities of artificial intelligence. The battle against cybercrime and the safeguarding of critical data can be greatly enhanced with continued developments in AGI and AI, supporting overall defense measures [13]. Since smart device networks and data are interconnected across cities and counties, it is challenging to determine which data is exposed over which pathways. To establish the hazards connected with each component, it is necessary to first identify the components of that technology [14]. Typical Cyber Threats and AGI-Enabled Defence Organizations and people alike are in serious danger from cyber threats, which are ever-evolving and increasingly sophisticated. On the other hand, the development of Artificial General Intelligence (AGI) might completely alter cyberwarfare tactics. By utilizing sophisticated machine learning algorithms and real-time threat intelligence, defence systems with artificial intelligence capabilities can successfully counteract common cyber-attacks. These systems are capable of independently identifying and retaliating against a wide range of assaults, such as DDoS attacks, phishing scams, and malware infestations. Proactive threat detection is made possible by AGI’s capacity to scan enormous volumes of data for patterns and analyze them, giving businesses an advantage over hackers. Furthermore, AGI can continually learn and adjust to novel assault vectors, strengthening defence measures’ resilience. Organizations may greatly improve their cyber defence capabilities and lessen the risks associated with common cyber-attacks by utilizing the power of artificial intelligence. These research findings will map each component to potential assaults. Knowing the issues that need to be addressed for each component and choosing the appropriate countermeasures will thus be helpful [3]. Standard Methods for Identifying Cyber Threats The purpose of this inquiry is to gain knowledge of cutting-edge methods and technologies that may be employed to aid with data security and transmission [15]. Security Issues in the Context of New Technologies Understanding the relevance of security issues with cutting-edge technology is the aim of this [16]. These queries were addressed using high-caliber publications from conferences and scholarly journals relevant to the research topic. Because technology is always evolving, new trends in the realm of Cyber Security are still in their early phases (Fig. 6.1). Background With the touch of a button, Modern men are capable of communicating with and receiving any type of data through email, voice, or pictures, but how safe is it that his data is being transferred or delivered to another person with no personal details being compromised? There’s a solution in Cyber Security. Networks, programs, and data are protected using a variety of techniques known as Cyber Security from harm, intrusion, and illegal access. Our personal information is not effectively protected by existing security measures due to the rapid growth of technology, which has increased the number of cybercrimes. As a result, this sector required top-notch security to ensure honest and effective transactions [17]. Because of this, Cyber Security has lately grown to be an issue. In addition to only safeguarding data in IT organizations, the scope of cyber security includes a variety of other industries as well, like cyber housing. Even the newest technologies, including Internet banking, online shopping, cloud services, smartphone use, etc., can cooperate with an elevated level of protection. Given that these developments contain essential data regarding one’s privacy, it has become a major worry [18]. For a country to be secure and to prosper economically, Cyber Security must be strengthened, and critical information infrastructure must be protected. The development of new government services and rules depends more and more on enhancing the security of online communication (and safeguarding users). A thorough and safer approach is needed to combat cybercrime. Any country’s safety and financial stability depend on bolstering cyber defenses and safeguarding valuable data infrastructure. Government regulations (and preserving online users) including the creation of modern services focus more and more on making the Internet safer. To tackle cybercrime, a detailed and safer method is required. Since technology measures alone cannot prevent any crime, social control agencies’ units of measurement must be permitted to investigate and penalize cybercrime. To prevent the loss of certain crucial knowledge, several nations and governments nowadays are measuring units and implementing rigorous rules on cyber security. To safeguard themselves against the increased number of crimes, everyone should get cyber security training [12] (Fig. 6.2). Motivation Cyber defence appears to be about to take a major turn thanks to artificial general intelligence (AGI). An appealing way to address the growing difficulties in protecting sensitive data and information is through artificial general intelligence (AGI), which is becoming more and more sophisticated and complicated every day. AGI is capable of detecting and responding to a variety of threats, including as DDoS attacks, phishing scams, and malware infections, on its own by utilizing cutting-edge machine learning algorithms and real-time threat data. Organizations may remain one step ahead of hackers thanks to proactive threat detection enabled by its capacity to analyze enormous volumes of data and spot trends. Furthermore, AGI’s ability to continually learn and adjust to novel methods of attack strengthens defence systems’ resilience, making it a useful tool for reducing the risks connected with cyber security. Organizations may greatly improve their cyber defence capabilities and guarantee the security of critical networks and client information by utilizing the power of artificial intelligence. Cyber Security entails several systems and gadgets connected by various technologies, and the amount of data these systems produce is substantial. If these systems and gadgets aren’t impervious, there might be serious harm and risk since user data and information that isn’t safeguarded could be used to incite riots and disobey the law [19]. As a result, I prefer to highlight in this article the most notable hazards connected to Cyber Security in technologies while also outlining deliberate defences against them. The survey paper is an empirical study that examines earlier research on Cyber Security trends in cutting-edge technology to manage the benefits and prospects of their use as well as the difficulties, dangers, and vulnerabilities of Cyber Security on technologies. Three milestones can be reached before the end of this semester and make up the research. The only restriction in preparing for this research is that the evaluated publications must have just recently been published if they are to span the time (2018–2022). Because risks to information security are a relatively new issue, there isn’t much research on the subject, which might be a restriction for this project. Research Contribution

#### Cyber-crime’s threat to businesses is on the brink – safe cyberspace is impossible without AGI

Arshad et al. 25 [Hira Arshad, Faculty of ICT at the Baluchistan University of Information Technology Engineering and Management Sciences, Ahthasham Sajid, researcher at the Department of Cyber Security and Data Science at Riphah Institute of Systems Engineering at Riphah International University, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Mehak Mushtaq Malik, researcher at the Department of Computer Science COMSATS University Islamabad, and Shahzad Latif, researcher at the Department of Computer Science at the Szabist University Islamabad Campus, 2025, “Chapter 6 A Survey on Cyber Security Encounters and AGI-Based Solutions,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Cyber Security Solutions Using AGI A more sophisticated type of artificial intelligence known as artificial general intelligence (AGI) is able to comprehend, acquire, and use information in a variety of fields. AGI has the potential to completely change how businesses safeguard their digital assets in the field of cyber security. Cyber security solutions may become more proactive, adaptable, and efficient by utilizing AGI’s capabilities, which include machine learning, pattern identification, and natural language processing. To discover potential risks and vulnerabilities, for example, AGI can analyze enormous volumes of data. This enables security teams to more efficiently allocate resources and prioritize their efforts. By automating routine processes like monitoring and handling security events, artificial intelligence (AGI) frees up human experts to work on more difficult problems. Organizations can benefit from enhanced security, lower operating costs, and more resistance to changing threats as a consequence. Conclusion The importance of cyber security is growing because of the world being more interconnected and networks becoming used to carrying out crucial activities. Cyber Security is a vast issue. With each passing year, criminality continues to develop whole new patterns, making knowledge protection more and more important. Using the most recent and revolutionary technology, Companies have problems in terms of not just safeguarding their system but also with newly formed services and knowledge to investigate. Furthermore the daily threats and new techniques are available for cyberspace. In order for us to live in a comfortable and secure cyberspace in the future, we must always make every effort to reduce cybercrimes, for which there is no suitable answer. This review article will look at the idea of cutting-edge technology, its advantages, and potential, as well as the problems, risks, and solutions related to Cyber Security. According to the studies we looked at, security is crucial to both our daily lives and upcoming advances. The study is complete. Since cyberattacks are one of the most urgent problems in modern technology, security is more important than ever to safeguard consumers’ privacy. This review on paper’s topics included how hackers might use weak spots in smart technology infrastructure as a breeding ground for data breaches and infrastructure failure. The development of Artificial General Intelligence (AGI) adds a new level of complexity to the current cyber security issues. AGI describes highly autonomous systems that perform better than humans in the majority of economically valuable tasks. It is essential to have proactive defense mechanisms and strong security protocols in order to handle the problems that artificial intelligence (AI) poses to cyber security. This entails making investments in R&D to keep ahead of new risks, encouraging industry-academia cooperation, and advocating for moral behavior in the development of AGI. International cooperation and legal frameworks are also necessary to guarantee the safe and responsible application of AGI technologies. Future Work Because databases are used to do useful tasks and the world is becoming more involved, computer protection may become a significant issue. Cybercrime starts to change as the millennium year progresses, first from new techniques and subsequently from data security. Companies find it challenging to protect their assets in the most current and chaotic systems since doing so require a variety of resources and expertise. This is due to the most recent security strategies as well as hazards that come to lightweight regularly. However, it appears that there is no effective cure for cybercrime. However, the proportionate reaction is insufficient. A global responsibility, Due to the ongoing and growing nature of cyber security concerns, duty, and deterrent systems must be built at the more dangerous frontier of cyber warfare. The results of this analysis demonstrated that cyber security adoption in cutting-edge technology is still in its infancy. Although there aren’t many excellent articles that discuss this combined terminology, there is a lot of interest in the topic. To advance smart technology services on smart devices, countries are now investing in cyber security. In the realm of cyber security, artificial general intelligence is a subject of great curiosity and worry. It recognizes that the development of AGI is inevitable and that there is ambiguity about how it will behave once it develops selfawareness. It is necessary to look into ways to reduce the possibility that AGI would turn unfriendly toward individuals, according to studies that indicate this is highly likely to happen. As a result, it is determined that controlling AGI is an important and pressing area of research. The paper focuses on creating a framework to define the essential components of upcoming containment technologies, despite the paucity of literature on containment tactics. Through the application of extant scholarly literature, the authors formulate a thorough ontology that identifies cyberspace, artificial general intelligence, and humans as necessary elements for successful containment. In addition, E-AntHocNet, CRAN plays an important role to secure communication within network [61–63].

#### AGI is better than narrow AI and humans at reducing cybercrime.

Varsha and Chakraborty 25 [Ar. Varsha, researcher at the Department of Forensic Science and Criminology at the Annai Fathima College of Arts and Science, and Pooja Chakraborty, member of the Chandigarh Police, 2025, “Chapter 15 Role of Artificial General Intelligence in the Prevention of Crime,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Artificial General Intelligence in Cybercrimes Living in a world that is digitalized, it is an understatement to seek privacy. Individuals themselves provide all their details to the network and expect privacy to be intact. This can never be cent percent assured considering the hacking and cyber malpractices that are being predominantly followed. To keep track of the illegal activities through social platforms and to prevent cybercrimes, AGI can be convenient. Countless crimes occur using the networking platform every day, and quite a few objectives to prohibit such actions includes • Improvised detection of cyber-threats and presenting a suitable solution to eradicate unwanted risks as well as to minimize the damage caused by any cyberattack. AI can identify these threats but sometimes disregard the issue due to technical issues or unsupervised scenarios. The decision-making skills of AI are not preferred as it doesn’t consider the situational factors and consequences but enacts the command. AGI on the contrary could act professionally and logically considering the factors to provide a better response. • Better performance on repetitive tasks to detect any suspicious entry breaching the security when the data is large. Human error while performing a repetitive task is often witnessed, but AGI can actively rectify those errors to provide an accurate and safe network. • Compared to humans, AGI has an understanding of historical data and thus can analyze the situation in an enhanced manner providing a clear adequate solution for any challenges opposed. • AGI can also provide security on multiple levels, ranging from endpoint security that analyzes user behavior to network security to tackle attacks and cloud security that controls and access permission to provide accurate decisions and avoid any malicious activities. • Any unnatural patterns or unethical illegal happenings can also be detected by AGI to prevent it from exploiting the users as well as the common people who could be affected by cyber terrorism or other violent acts. On the whole, cyber-threats and cyber-attacks can only be prevented by constant monitoring of the systems and their network, building up a higher level of security, and keeping track of every action that is being done by every user of a network. AGI can keep monitoring the system and inform the user or head of the network in any cases of suspicious activities or look up on itself to decide how to proceed and react to the situation [14]. A merit of AGI over AI is that assistance is always required for AI either by programs or by humans which is not the case for AGI as it self-learns and performs a task. Artificial General Intelligence in Offline Crimes

#### AGI solves crime – enhanced surveillance, data analysis, detective assistance, and predictive policing

Varsha and Chakraborty 25 [Ar. Varsha, researcher at the Department of Forensic Science and Criminology at the Annai Fathima College of Arts and Science, and Pooja Chakraborty, member of the Chandigarh Police, 2025, “Chapter 15 Role of Artificial General Intelligence in the Prevention of Crime,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Artificial General Intelligence in Offline Crimes Crimes that occur daily around us without the influence of technology like murder, homicide, destruction of property, and so on can also be prevented by using the AGI model. Situations and scenes where the AGI can be put in an application to prevent the crime can be as follows: • Surveillance cameras are prevalent all over the place, but infusing these with AGI can stimulate the result of crime prevention by – Ensuring whether any violations of human rights are happening anywhere to immediately alert the concerned officials nearby. – Psychology assessment by cameras and AGI that could detect people’s mentality of those who wish to commit suicide or self-harm themselves and ensure help is reached out to these individuals. – Software that could recognize facial features to detect any intention of an individual who tries to commit a crime anywhere by any means or even keep track of someone’s criminal history based on facial recognition to alert the officials of the offender. – Reporting suspicious activities based on machine learning which can sometimes prevent real crimes. • Not limiting the above-mentioned to surveillance cameras, various other means like monitored drones could also be employed for the same. • Police officials can also use mind reading or mental assessment software which could be beneficial both during criminal interrogation as well as understanding the criminal profile of an offender. • Artificial Neural Networks are a kind of machine learning through which the computer is programmed to function similarly to the human neural system which helps to understand and interpret images and speech which can aid in criminal investigation. • AGI with proper assistance can investigate, interrogate, and identify a criminal based on input algorithms and facts of the crime discovered. With repeated exposure, AGI itself can succeed these in comparatively less time, and with more accuracy than a normal human being would. • Imports and exports can use AGI to detect the smuggling of illicit drugs and valuable articles illegally thus preventing inter-country and intra-country smuggling and trafficking [15]. • The pharmaceutical sector can employ AGI to detect unusual purchases of any dose beyond the normal range which can be used for murder or suicide to prevent them. • Not limited to petty crimes, AGI can also help to prevent terrorism and war based on its predictive analysis and strategy planning to combat against the enemy ensuring the restoration of national peace. The particulars are a few easy conventional applications of AGI to prevent and reduce crimes that occur regularly around us. Not limiting the innovation for these, there are yet so many sectors and scenarios where AGI can play the role of a game changer in controlling crimes. Predictive Analysis of AGI and Crime Prevention Crime prevention can be performed only when the people, geographical area, and even the networks are under constant supervision. Considering the vast residential area, the population of every state, and the numerous gadgets that are being owned by every single individual, it is indeed a tedious process to keep track of everyone and everything. To create a solution for this particular problem, it is essential to keep tabs on everyone but focus on a particular area and people where the possibility of crime commission can be high [16]. Obtaining a high target area or specific target individual is done using the predictive analysis of AGI. Plenty of data are simultaneously analyzed by the system to get insight into crime hotspots all over the world. This process of analyzing data to identify target areas to bring to a conclusion where more attention is required is termed the Predictive Policing of AGI. Historic data of location, and crime history of any area and every individual are given as input to receive statistics of crime forecast shortly alerting where extra attention is required to prevent crimes from occurring. The algorithms have proven to be successive so far in Artificial Intelligence and can even be much upgraded in Artificial General Intelligence for its additional features and intellectual characteristics. AI can predict the probability of crime but AGI can also decide how the crime can be approached to prevent it without any harm to anybody. Various models of predictions are being made by the AGI and the primarily applied predictive analysis are as follows. Geographic hotspots Certain geographic areas are prone to the crime rate as well as the existence of criminal gangs. When existing criminal rates are being enrolled into the machine, the likeliness of crime to be committed in a particular area can be predicted and intimated. Based on the hotspots obtained by the analysis of AGI, officials can likewise segregate their importance of supervision as per the requirements. This technology is especially helpful for tourists, who may be unaware of the new places. AGI has advanced to a level where even the type of crime a place is prone to can be highlighted. Manner of crime Not limited to the specific areas that are popular to crime, AGI can analyze the pattern of crime that is prevalent at a location. For instance, an area and its people can be said to deal with drugs and smuggling. So, while supervising these areas, the officials can focus on the import and export goods more. Thus, identifying the trendsetting crime of any location can help to comprehend when and what should be focused on by the officials to restrict the happenings of the crime. Utilization of resource Identifying hotspots and crime trends can further pave the way on how the available resources of the government must be employed to prevent crimes. Understanding the points of entry and exit of an area and commonly preferred routes by the criminals can help the police to give extra attention to those places. The importance of the types of resources required to monitor and the intensity of requirements can be planned and executed by the AGI. Offender recognition Keeping track of offenders and special attention to criminals who are prone to repeat their offenses are important as these people need extra care and attention to prohibit them from committing crimes. These offenders are to be considered high-risk individuals who can have greater potential in crime execution. AGI based on facial recognition can constantly keep these individuals on track to prevent the increase in the crime rate. These predictions can boost AGI-powered law enforcement to make dynamic decisions and generate the necessary preventive measures in suitable areas at suitable times. Optimizing the various resources of law that include the budget and the labor force is another key essential advantage of using AGI in the planning. With time, AGI also modifies its data thus providing real-time perception and awareness both to the society as well as the government officials. The role of AGI is never restricted to identification, monitoring, and strategy planning; it is also one step ahead in taking last-minute critical decisions that are acceptable and authenticate. AGI in Crime Scene Detection

#### AGI-aided investigations solve crime better than humans

Varsha and Chakraborty 25 [Ar. Varsha, researcher at the Department of Forensic Science and Criminology at the Annai Fathima College of Arts and Science, and Pooja Chakraborty, member of the Chandigarh Police, 2025, “Chapter 15 Role of Artificial General Intelligence in the Prevention of Crime,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

AGI in Crime Scene Detection Crime prevention never implies stopping a crime from happening; it could also indicate breaking a series of crime patterns and preventing further occurrence of crimes. In instances of serial killing, where a similar modus operandi (method of operation) is carried out, AGI can analyze various pieces of evidence found at the crime scene and during the investigation to identify the killer [17]. Application of AGI in crime scenes can be witnessed in the possible ways which could include the following: • Analysis of multiple crime scene photography and videography to deduct minute trace pieces of evidence and also any abnormalities in the scene that cannot be deducted or noticed by investigators. • Comparing the anomaly pieces of evidence to their origin by comparing them to the databases to get an idea of what a particular piece of evidence is and what role it could have played at the scene of the crime. • Vulnerabilities of human beings in recognizing a specific individual or piece of evidence in video or images from CCTV cameras can be avoided when AGI is employed. • Tampering of visual or audio evidence is negligible by human beings, but these machines can even identify a minute change or alteration in them alerting the officials. • DNA evidence is one of the most significant vital pieces of evidence in criminal evidence, and usually it can be tampered with during collection and transportation. Researchers assisted with AGI algorithms can deduce particular required samples from the collection of various DNA and also the database of AGI can easily give identity to the individual to whom the DNA sample belongs. • Firearm identification is a developing aspect using this artificial intelligence whereby the images of the injury and medical reports of the injured persons are acknowledged by AGI. Based on these, inferences are made regarding the type of firearm, model, and type of bullet used. Sometimes, even the user can be deduced based on the model of firearm used. • Since AGI also could help in the field of medicine, it can also help to access the cause and manner of death of a person based on the autopsy report which could have been missed by the experts. Analyzing the crime scene as well as the pieces of evidence obtained, the AGI can link the happenings of a crime. Perpetrator identification is also at the fingertips with the improvised version of artificial intelligence like never before. So, the further commission of crimes by these people can be also prevented thus restricting the repeated offender scenario. Development of Artificial General Intelligence

#### AGI is key to sustainable development – grid management, urban planning, and traffic management.

Singh and Kaunert 25 [Bhupinder Singh, researcher at Sharda University, and Christian Kaunert, Professor of International Security at Dublin City University with a PhD in International Politics & an MSc in European Politics from the University of Wales Aberystwyth, a BA (Hons) European Business from Dublin City University, ESB Reutlingen and a BA (Hons) Open University, 2025, “Chapter 12 Dynamic Landscape of Artificial General Intelligence (AGI) for Advancing Renewable Energy in Urban Environments: Synergies with SDG 11—Sustainable Cities and Communities Lensing Policy and Governance Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Introduction Artificial General Intelligence (AGI) represents a paradigm shift in the field of artificial intelligence, striving to imbue machines with human-like cognitive capabilities that extend beyond narrow task-specific abilities. AGI, as opposed to conventional limited AI, seeks to emulate the breadth of human intellect, allowing machines to perceive, learn, reason, and adapt across multiple areas. AGI’s promise resides in its ability to address complex and varied problems, making it a formidable force in the pursuit of creating renewable energy solutions within the context of Sustainable Development Goal 11 (SDG 11)—Sustainable Cities and Communities. AGI goes beyond the limits of specialized AI by attempting to imitate human cognitive flexibility. AGI’s fundamental goal is to reach human-level intelligence, allowing robots to perform a wide range of intellectual tasks with the same adaptability and comprehension as humans [1]. AGI’s capabilities are broad and powerful, backed up by its capacity to grasp complicated concepts, learn from experiences, and be creative. AGI systems grow over time through adaptive learning and optimization, continuously increasing their performance based on acquired data. AGI’s strength in renewable energy is its ability to optimize energy consumption patterns in urban environments, adjusting to changing demands and conditions [2]. Furthermore, AGI excels at complicated problem-solving. It can deconstruct complex issues, assess various variables, and generate novel solutions. AGI has the ability to address the multiple difficulties of energy distribution, demand forecasting, and **grid** **stability** in the context of sustainable cities, which necessitate detailed analysis and innovative techniques. The other key strength of AGI is data integration and analysis. It excels in managing and processing massive datasets from various sources, identifying patterns, and extracting valuable insights. This skill becomes useful when it comes to processing data from renewable energy sources, weather forecasts, and urban demography in order to influence efficient energy usage, distribution, and infrastructure development. The capabilities of AGI extend to decision-making under uncertainty. It is capable of making educated decisions even when information is sparse or unclear. This ability is especially important in optimizing energy storage and distribution within the volatile urban landscape, where a plethora of factors can influence energy supply and demand. AGI’s cross-domain proficiency allows it to operate outside of certain knowledge domains. It can pull lessons from numerous domains and apply them imaginatively to a variety of situations. In the context of SDG 11, AGI has the potential to combine urban planning concepts with novel renewable energy techniques, enabling a comprehensive and harmonious approach to sustainable urban development. The incorporation of AGI within the scope of SDG 11 opens up previously unseen prospects for the creation of smart and sustainable urban settings [3, 4]. The capabilities of AGI are well aligned with the goals of SDG 11, contributing to energy efficiency, renewable energy adoption, and resilient urban infrastructure. AGI has the potential to alter the way energy is utilized within buildings, street lights, and **transportation** networks in the field of energy efficiency and management. Renewable energy integration is a critical component of long-term urban development. AGI-powered solutions can forecast renewable energy generation patterns and dynamically balance them with consumption, assisting in the development of reliable and stable energy grids. The flexibility of AGI to adapt to changing situations guarantees that renewable resources are used efficiently. AGI has the potential to alter urban mobility and planning. AGI can alleviate traffic congestion, reduce energy consumption, and improve the overall urban experience by optimizing traffic flow and public transportation routes. This, in turn, is consistent with the goal of constructing sustainable and habitable cities. Resilient infrastructure is fundamental for urban sustainability, especially in the face of challenges such as **climate change** and natural disasters. AGI’s capacity to analyze diverse datasets can help identify vulnerabilities within urban infrastructure and develop strategies to enhance resilience. By contributing to the creation of adaptive and responsive urban systems, AGI aids in building cities that are prepared for various contingencies [5]. While AGI offers remarkable potential, its deployment must be accompanied by a deep commitment to ethical considerations and responsible development. Issues of bias, transparency, and accountability must be addressed to ensure that the benefits of AGI are equitably distributed among diverse populations. Ensuring that AGI systems are developed and deployed in a manner that upholds ethical principles is essential to prevent unintended negative consequences, particularly in the context of complex urban environments. AGI stands as a beacon of innovation and transformative potential. Its capabilities in addressing intricate challenges, learning from experiences, and facilitating cross-domain insights position it as a formidable tool in the advancement of renewable energy solutions within the framework of SDG 11. By harnessing AGI’s cognitive abilities, we can pave the way for sustainable cities and communities that thrive on efficient energy usage, renewable sources, and resilient urban planning. However, this journey must be guided by ethical considerations and responsible practices to ensure that AGI serves as a force for positive change in the quest for sustainable development [4]. Objectives of the Chapter

#### AGI energy management solves grid stability and integrates renewables, which solves blackouts and climate change

Singh and Kaunert 25 [Bhupinder Singh, researcher at Sharda University, and Christian Kaunert, Professor of International Security at Dublin City University with a PhD in International Politics & an MSc in European Politics from the University of Wales Aberystwyth, a BA (Hons) European Business from Dublin City University, ESB Reutlingen and a BA (Hons) Open University, 2025, “Chapter 12 Dynamic Landscape of Artificial General Intelligence (AGI) for Advancing Renewable Energy in Urban Environments: Synergies with SDG 11—Sustainable Cities and Communities Lensing Policy and Governance Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

This understanding enables educated urban planning decisions, resulting in the development of energy-efficient buildings, sustainable land use, and resilient infrastructure. The predictive powers of AGI aid in identifying weaknesses, enabling proactive disaster management solutions that line with SDG 11’s resilience objectives. AGI applications in sustainable cities provide enormous potential for realizing SDG 11’s vision of sustainable urban development. AGI-driven innovations enable cities to embrace technology for holistic advancement, from minimizing energy use and integrating renewable energy to improving urban mobility and resilient infrastructure [12]. It paves the path for cities that are not just intelligent, but also ecologically responsible, socially inclusive, and resilient in the face of the complex challenges posed by urbanization and climate change by using AGI’s capabilities. Smart Energy Management The intersection of Smart Energy Management and Artificial General Intelligence (AGI) is a game changer in the field of sustainable urban development. AGI’s cognitive capability is a formidable ally in Smart Energy Management, which is defined as the intelligent optimization of energy production, distribution, and consumption. The ability of artificial intelligence (AGI) to perceive complex patterns, process massive datasets, and adjust in real-time enables cities to raise their energy systems to unprecedented levels of efficiency, resilience, and sustainability. The implementation of AGI in smart energy management limits its predictive analytical capacity. AGI can predict energy demand patterns with surprising accuracy by taking into account elements like weather, previous consumption habits, and even future events [13]. This predictive capability enables energy grids to alter energy generation and distribution proactively, reducing waste and assuring a seamless balance of supply and demand. AGI-equipped systems promote effective load distribution by responding to dynamic changes in consumption, decreasing strain on the grid during peak hours, and mitigating the likelihood of outages. Because of AGI’s adaptive learning capabilities, it can continuously improve energy management tactics over time. AGI systems fine-tune their algorithms as data accumulates, improving their accuracy in forecasting demand patterns and optimizing energy distribution. The integration of renewable energy sources is one of the pillars of smart energy management. The function of AGI in this integration is revolutionary. AGI can estimate renewable energy generation patterns by taking elements such as sunshine intensity, wind speed, and water flow into account. AGI enables cities to easily integrate solar, wind, and hydropower into their energy mix by merging this data with real-time demand estimates. This not only minimizes dependency on fossil fuels, but also improves electricity grid stability by effectively leveraging the fluctuation of renewable resources. The ability of AGI to make real-time decisions is critical in improving grid resilience and stability. During unanticipated events such as abrupt increases in demand or outages in energy supply, AGI can quickly adjust and optimize energy distribution to avoid cascading failures. This agility contributes to the reliability of urban energy systems, safeguarding against energy shortages and blackouts [14]. However, AGI’s potential in smart energy management extends beyond mere technical optimization. It also ushers in a new era of consumer engagement and empowerment. AGI-powered systems can provide real-time insights into energy consumption, allowing consumers to make informed decisions about their usage patterns. By facilitating energy conservation at the individual level, AGI contributes to the broader sustainability goals of cities, as outlined in Sustainable Development Goal 11. The synergy between Smart Energy Management and Artificial General Intelligence holds immense promise for redefining urban energy landscapes. By harnessing AGI’s predictive and adaptive capabilities, cities can optimize energy distribution, seamlessly integrate renewable sources, and enhance grid resilience. The result is a transformation toward energy systems that not only meet the demands of urbanization, but also contribute to sustainable development, environmental conservation, and improved quality of life for citizens in the cities of the future [15]. Real-time Energy Demand Prediction and Optimization

#### AGI solves traffic, massively reducing carbon emissions.

Singh and Kaunert 25 [Bhupinder Singh, researcher at Sharda University, and Christian Kaunert, Professor of International Security at Dublin City University with a PhD in International Politics & an MSc in European Politics from the University of Wales Aberystwyth, a BA (Hons) European Business from Dublin City University, ESB Reutlingen and a BA (Hons) Open University, 2025, “Chapter 12 Dynamic Landscape of Artificial General Intelligence (AGI) for Advancing Renewable Energy in Urban Environments: Synergies with SDG 11—Sustainable Cities and Communities Lensing Policy and Governance Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Urban Mobility and Transportation: Transforming Urban Life with AGI The rapid pace of urbanization has brought to the forefront the urgent need for reimagining urban mobility and transportation systems. As cities become bustling centers of activity, challenges such as traffic congestion, air pollution, and inefficient transportation networks have intensified. The incorporation of Artificial General Intelligence (AGI) appears as a disruptive force capable of altering how individuals travel through metropolitan landscapes. The cognitive talents, data processing prowess, and adaptive decision-making capabilities of AGI have the potential to develop intelligent, efficient, and sustainable urban mobility ecosystems [26]. The AGI’s utility in urban transportation stems from its ability to interpret massive volumes of real-time data from a variety of sources, including traffic sensors and GPS devices, as well as weather forecasts and social media platforms. AGI may evaluate traffic trends, predict congestion points, and dynamically optimize transportation routes using this data-driven method. As a result, commuters enjoy shorter travel times, less fuel consumption, and lower emissions, resulting in a more smooth and delightful urban commuting experience. The role of AGI’s in urban mobility aligns seamlessly with the objectives of the Sustainable Development Goal—SDG 11. By facilitating more efficient transportation, AGI contributes to reduced carbon emissions, improved air quality, and enhanced overall urban livability. The optimization of transportation systems resonates with the goal of providing access to safe, affordable, and sustainable transportation for all residents, ensuring that urban mobility is inclusive and equitable. By analyzing transportation data, AGI can offer insights into areas of high demand, aiding in the strategic placement of public transportation stops and bike-sharing stations [27]. This strategic urban planning encourages the use of alternative modes of transportation, reducing reliance on private vehicles and fostering environmentally conscious mobility choices. However, the full realization of AGI’s potential in urban mobility requires addressing challenges such as data privacy, cybersecurity, and ethical considerations. Collaborations between AI experts, transportation planners, and policymakers are pivotal in developing responsible frameworks that prioritize the safety, security, and well-being of urban residents. As AGI transforms urban travel from a source of congestion to a driver of efficiency and environmental stewardship, it sets the stage for urban environments where mobility enhances the quality of life for all citizens while embracing the principles of sustainability and progress [28].

Energy Efficiency in Buildings: Fostering Sustainable Urban Environments and Aligning with SDG 11

#### AGI enables smart grids that optimizes energy usage based, removing

Singh and Kaunert 25 [Bhupinder Singh, researcher at Sharda University, and Christian Kaunert, Professor of International Security at Dublin City University with a PhD in International Politics & an MSc in European Politics from the University of Wales Aberystwyth, a BA (Hons) European Business from Dublin City University, ESB Reutlingen and a BA (Hons) Open University, 2025, “Chapter 12 Dynamic Landscape of Artificial General Intelligence (AGI) for Advancing Renewable Energy in Urban Environments: Synergies with SDG 11—Sustainable Cities and Communities Lensing Policy and Governance Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Energy Efficiency in Buildings: Fostering Sustainable Urban Environments and Aligning with SDG 11 In the global pursuit of Sustainable Development Goal 11 (SDG 11)—Sustainable Cities and Communities, energy efficiency in buildings emerges as a pivotal pathway toward creating environmentally conscious, resilient, and livable urban spaces. The built environment is a major consumer of energy and contributor to greenhouse gas emissions, making it imperative to prioritize energy-efficient practices. Energy efficiency in buildings refers to a variety of measures aimed at lowering energy use while preserving or improving comfort and functionality [29]. These solutions include enhancing insulation, optimizing heating, cooling, and lighting systems, deploying energy-efficient products, and implementing intelligent building management systems. Implementing these measures leads to lower energy expenses for tenants, lower carbon emissions, and less load on energy infrastructure. This, in turn, contributes to the broader goals of SDG 11, which envisions cities that are environmentally conscious, inclusive, and resilient. By embracing energy efficiency, cities can create an intricate web of synergies with various targets outlined within SDG 11. The notable intersection lies in the goal of providing affordable and clean energy (SDG 7). Energy-efficient buildings consume less energy, translating into lower utility bills for residents and businesses. This not only enhances affordability, but also frees up resources that can be redirected toward other essential needs, ultimately contributing to the creation of economically vibrant and socially inclusive urban environments. The essence of energy efficiency in buildings also aligns with the aspiration to create sustainable transportation systems (SDG 11.2). Reduced energy consumption within buildings means less strain on energy grids, which in turn lowers overall demand for energy generation. This decrease in energy demand indirectly reduces the need for fossil-fuel-powered transportation to supply energy, resulting in lowered carbon emissions and improved air quality. Furthermore, energy-efficient buildings often incorporate sustainable transportation options such as bike storage, electric vehicle charging stations, and proximity to public transit, encouraging environmentally friendly mobility choices [30]. The integration of energy-efficient practices resonates with SDG 11 which emphasizes access to safe, affordable, and sustainable housing for all. Energy-efficient buildings offer improved indoor air quality, thermal comfort, and noise reduction, contributing to healthier and more comfortable living spaces. The reduction in energy bills associated with energy-efficient designs benefits marginalized communities by alleviating energy poverty and enabling more equitable access to basic needs [31]. While energy efficiency in buildings holds great promise, its realization does face challenges such as initial costs, retrofitting existing structures, and policy implementation. However, the long-term benefits far outweigh these challenges. The adoption of energy-efficient technologies and practices not only leads to environmental conservation, but also stimulates local economies through job creation in construction, technology installation, and maintenance sectors [32]. AGI-enabled Building Energy Management Systems: Revolutionizing Sustainable Urban Infrastructure The fusion of Artificial General Intelligence (AGI) with building energy management systems heralds a new era in the realm of sustainable urban development. As cities battle with issues such as energy consumption, environmental impact, and efficient resource management, AGI emerges as a revolutionary force capable of transforming buildings into intelligent, adaptable, and energy-efficient entities. AGIenabled building energy management systems represent a paradigm shift in how we optimize energy usage, effortlessly harmonizing with sustainability and environmental preservation aims [33]. These systems can monitor real-time energy use patterns as well as evaluate complicated data streams such as weather forecasts, occupancy trends, and building usage patterns. AGI systems can anticipate energy demand variations, optimize heating, cooling, and lighting systems, and dynamically alter energy use based on real-time conditions by assimilating and understanding this data. This predictability leads to significant energy savings, low operational costs, and less environmental impact. These systems can learn from previous data on energy consumption and user behavior, changing over time to optimize energy consumption patterns. For example, an AGI-enabled system can identify patterns of energy wastage and automatically adjust building systems to prevent unnecessary consumption. This continuous learning process fine-tunes energy management strategies, ensuring that buildings operate at peak efficiency under diverse conditions [34]. In the context of smart cities, where various components of urban infrastructure communicate and collaborate, AGI acts as an orchestrator. These systems can interact with broader city-wide data streams, such as transportation and weather data, to optimize energy usage in response to larger urban dynamics. For instance, during periods of peak energy demand, an AGI-enabled building energy management system could receive signals from the energy grid and automatically adjust building energy consumption to reduce strain on the system. The optimization of energy consumption within buildings directly resonates with SDG 11’s aim of enhancing resource efficiency and creating resilient urban environments. However, challenges such as data security, interoperability, and potential biases in AI algorithms must be addressed to fully leverage the potential of AGI-enabled building energy management systems. Collaborative efforts among AI experts, building professionals, and policymakers are essential in ensuring responsible and equitable deployment. As these systems become integral components of future urban infrastructure, they propel cities toward the realization of a sustainable, resilient, and energy-efficient future, in alignment with the principles of SDG 11 and the broader vision of a more sustainable world. Promoting Community-level Renewable Energy Projects Through AGI-assisted Planning

#### AGI builds community buy-in with custom-tailored solutions specific to their needs

Singh and Kaunert 25 [Bhupinder Singh, researcher at Sharda University, and Christian Kaunert, Professor of International Security at Dublin City University with a PhD in International Politics & an MSc in European Politics from the University of Wales Aberystwyth, a BA (Hons) European Business from Dublin City University, ESB Reutlingen and a BA (Hons) Open University, 2025, “Chapter 12 Dynamic Landscape of Artificial General Intelligence (AGI) for Advancing Renewable Energy in Urban Environments: Synergies with SDG 11—Sustainable Cities and Communities Lensing Policy and Governance Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

Promoting Community-level Renewable Energy Projects Through AGI-assisted Planning The combination of Artificial General Intelligence (AGI) with planning and implementation of community-level renewable energy projects represents a groundbreaking approach to advancing sustainable development. As the world grapples with the urgent need to transition toward cleaner energy sources, community-driven initiatives play a pivotal role in shaping the future of energy landscapes. AGI’s cognitive capabilities, data analysis prowess, and adaptive decision-making offer unprecedented potential to foster effective and efficient community-level renewable energy projects, aligning seamlessly with the goals of environmental conservation and inclusive development [35]. AGI-assisted preparation in community-level renewable energy plans begins with the collection of data from several sources, such as topographical, weather, and demographic data. AGI can model energy consumption patterns, locate viable renewable energy sources (such as solar, wind, or biomass), and estimate energy demand trends within the community by processing these inputs. This data-driven method enables planners and stakeholders to make informed decisions about the best sorts of renewable energy systems to adopt, their capacities, and how they should be distributed throughout the community. By simulating different configurations, AGI helps identify the most efficient and cost-effective pathways for implementing renewable energy systems tailored to the specific needs of the community. This optimization process fosters resilience, ensuring that renewable energy projects contribute to long-term sustainability. The AGI’s adaptability allows for dynamic adjustments as community dynamics evolve. Changes in population, energy demand, and technological advancements can be accommodated in the planning process. AGI continuously refines its predictions and recommendations based on real-time data, facilitating energy systems that remain responsive to changing conditions and are adaptable to unforeseen challenges [36]. The synergy between AGI-assisted planning and Sustainable Development Goal 11 (SDG 11)—Sustainable Cities and Communities—is evident. By empowering communities to implement renewable energy projects, AGI contributes to accessible and sustainable energy access (SDG 7) for all residents. The locally-focused nature of community-level projects aligns with SDG 11’s objective of creating inclusive, safe, and resilient urban environments that prioritize the well-being of their inhabitants. Through transparent data-driven insights, AGI fosters informed discussions and consensus-building, ensuring that the projects reflect the unique aspirations and priorities of the community. This engagement resonates with SDG 11.3, which seeks to provide affordable housing and basic services to all residents while promoting inclusivity and participation [37]. While AGI offers transformative potential, challenges such as data privacy, ethical considerations, and community engagement must be addressed for responsible and equitable implementation. Collaborations among AI experts, energy professionals, urban planners, and community representatives are essential in harnessing AGI’s potential to drive sustainable development through community-level renewable energy projects. Reduced Carbon Emissions Through Optimized Energy Consumption: The Role of Technological Innovation As the global community confronts the pressing challenges of climate change and environmental degradation, optimizing energy consumption emerges as a powerful strategy to reduce carbon emissions and mitigate the impacts of fossil fuel-driven activities. Technological innovation, driven by advancements in Artificial General Intelligence (AGI), offers a transformative avenue to achieve this goal. AGI’s predictive capabilities, real-time adaptability, and data-driven insights hold the potential to revolutionize how energy is consumed, resulting in significant reductions in carbon emissions and a more sustainable future. By processing information from smart meters, sensors, weather forecasts, and historical energy usage patterns, AGI can develop highly accurate models of energy demand and consumption behavior. This predictive power enables AGI to forecast peak energy demand periods, thus empowering energy grids to adjust energy generation and distribution accordingly. By intelligently matching supply with demand, AGI reduces the need for excess energy production, minimizing the burning of fossil fuels and subsequently cutting down on carbon emissions [38]. Through continuous data analysis, AGI can identify patterns of wasteful energy usage and make instant recommendations for energy-saving measures. This might involve optimizing lighting systems, adjusting thermostat settings, or even suggesting optimal times for using energy-intensive appliances. Such interventions, based on real-time data, contribute to immediate reductions in energy consumption and the associated carbon footprint. AGI can dynamically manage the distribution of energy across the grid, minimizing energy losses during transmission and ensuring that energy is delivered precisely where and when it’s needed. In transportation, AGI can optimize traffic flow, route planning, and vehicle charging, reducing congestion and idling, and consequently, cutting down on vehicle emissions [39]. The alignment between AGI-enabled optimized energy consumption and environmental sustainability resonates profoundly with global sustainability initiatives, including Sustainable Development Goal 13 (SDG 13)—Climate Action. By mitigating carbon emissions through more efficient energy use, AGI contributes directly to SDG 13’s objective of combating climate change and its impacts. The collaboration between technological innovation and environmental stewardship showcases a harmonious balance between human progress and planetary health. However, realizing the full potential of AGI-driven energy optimization requires overcoming challenges such as data privacy, cybersecurity, and integration with existing infrastructure [40]. Moreover, ethical considerations in AI algorithm design must ensure equitable benefits across diverse communities. Ethical and Societal Implications of AGI in Sustainable Cities: Navigating a Complex Landscape

#### AGI stops crop diseases, increasing food yields and agriculture resilience

Asif et al. 25 [Ali Asif, researcher at the National University of Sciences and Technology, Hassan Asif, researcher at the Pakistan Institute of Engineering and Applied Sciences, Abdullah Akbar, researcher at the National University of Computer and Emerging Sciences, Maqsood M. Khan, researcher, Shahzad Latif, researcher at the Department of Computer Science at Szabist University, Muhammad Ameer Hamza, researcher at the Department of Civil Engineering at COMSATS University Islamabad, and Abdur Rehman Khan, researcher at the Air University in Pakistan, 2025, “ Chapter 16 AGI-Enabled Robotics for Healthcare Industry,” Artificial General Intelligence (AGI) Security, https://www.springerprofessional.de/en/artificial-general-intelligence-agi-security/27634942]/Kankee

AGI for Early Disease Identification Advancements in artificial intelligence (AGI) have revolutionized the field of health care. One important application of this is the early identification of disease, which allows for timely treatment and handling of the issue. Traditional methods of disease detection often relied on symptom-based diagnoses or routine screenings, which can be limited in their accuracy and efficiency. Furthermore, sometimes the disease is only noticed when the damage has already occurred. AGI-driven algorithms can analyze changes in a patient’s vital signs over time to detect early signs of conditions such as heart disease or diabetes, allowing for timely interventions and prevention strategies. These models are trained on a very large amount of dataset. This model is then able to detect subtle patterns and anomalies in medical data, allowing for the identification of diseases at their earliest stages. Recently, many ML models have been created for early diagnosis of COVID-19, cancer, and diabetes [46]. Further studies are also being done using methods like supervised learning and deep learning to develop better algorithms for early disease identification. Techniques like the Boltzmann machine, K nearest neighbor (kNN), support vector machine (SVM), decision tree, logistic regression, fuzzy logic, and artificial neural network are being used for early diagnosis of disease [47]. Cancer is the second leading cause of death in the world [48]. It is an aggressive disease, and early diagnosis of it is extremely important to enhance a patient’s survival rate. AGI systems, after decades of research and experimentation, are now being used to assist the healthcare processes. Artificial Intelligence-based models, which are trained on a very large dataset, are good at recognizing patterns and can thus assist in the early identification of genomic variants and abnormal protein communications [49]. As described by Borhani and Borhani [50], different types of machine learning models have been used for the detection of early-stage tumors to diagnose and prevent Bladder cancer. Not only has AGI revolutionized healthcare optimization in humans, but it is also being used to monitor and improve the health of different types of **crops** as well. AGI systems have recently seen applications in the field of agriculture, where they are helping in the detection and control of illness at the early stage of plant life, thus optimizing the conditions for its healthy growth [51]. Similarly, deep learning models are being deployed for the timely identification and control of diseases in apples [52]. Such models have the potential to save **huge amounts of revenue** for the agriculture industry as losses due to plant diseases are detected and controlled at an earlier stage. In conclusion, the remarkable advancements in artificial intelligence have ushered in a transformative era for health care, offering invaluable tools for the early detection and diagnosis of diseases. As we continue to witness the evolution of AGI in health care and agriculture, it is evident that its impact on early disease identification and prevention will be instrumental in shaping a healthier and more resilient future for both humans and crops alike. Assisting Patients Through AGI-Based Chatbots

### Contention 2: Economy

#### Superintelligence innovation drives hypergrowth – automates labor and accelerates R&D

Aschenbrenner 24 [Leopold Aschenbrenner, research affiliate with the Oxford Global Priorities Institute and researcher on superalignment at OpenAI with a BA in Mathematics-Statistics and Economics from Columbia University, 06-2024, “II. From AGI to Superintelligence: the Intelligence Explosion,” Situational Awareness, https://situational-awareness.ai/from-agi-to-superintelligence/]/Kankee

AI progress won’t stop at human-level. After initially learning from the best human games, AlphaGo started playing against itself—and it quickly became superhuman, playing extremely creative and complex moves that a human would never have come up with. We discussed the path to AGI in the previous piece. Once we get AGI, we’ll turn the crank one more time—or two or three more times—and AI systems will become superhuman—vastly superhuman. They will become qualitatively smarter than you or I, much smarter, perhaps similar to how you or I are qualitatively smarter than an elementary schooler. The jump to superintelligence would be wild enough at the current rapid but continuous rate of AI progress (if we could make the jump to AGI in 4 years from GPT-4, what might another 4 or 8 years after that bring?). But it could be much faster than that, if AGI automates AI research itself. Once we get AGI, we won’t just have one AGI. I’ll walk through the numbers later, but: given inference GPU fleets by then, we’ll likely be able to run many millions of them (perhaps 100 million human-equivalents, and soon after at 10x+ human speed). Even if they can’t yet walk around the office or make coffee, they will be able to do ML research on a computer. Rather than a few hundred researchers and engineers at a leading AI lab, we’d have more than 100,000x that—furiously working on algorithmic breakthroughs, day and night. Yes, recursive self-improvement, but no sci-fi required; they would need only to accelerate the existing trendlines of algorithmic progress (currently at ~0.5 OOMs/year). Automated AI research could probably compress a human-decade of algorithmic progress into less than a year (and that seems conservative). That’d be 5+ OOMs, another GPT-2-to-GPT-4-sized jump, on top of AGI—a qualitative jump like that from a preschooler to a smart high schooler, on top of AI systems already as smart as expert AI researchers/engineers. There are several plausible bottlenecks—including limited compute for experiments, complementarities with humans, and algorithmic progress becoming harder—which I’ll address, but none seem sufficient to definitively slow things down. Before we know it, we would have superintelligence on our hands—AI systems vastly smarter than humans, capable of novel, creative, complicated behavior we couldn’t even begin to understand—perhaps even a small civilization of billions of them. Their power would be vast, too. Applying superintelligence to R&D in other fields, explosive progress would broaden from just ML research; soon they’d solve robotics, make dramatic leaps across other fields of science and technology within years, and an industrial explosion would follow. Superintelligence would likely provide a decisive military advantage, and unfold untold powers of destruction. We will be faced with one of the most intense and volatile moments of human history. Automating AI research We don’t need to automate everything—just AI research. A common objection to transformative impacts of AGI is that it will be hard for AI to do everything. Look at robotics, for instance, doubters say; that will be a gnarly problem, even if AI is cognitively at the levels of PhDs. Or take automating biology R&D, which might require lots of physical lab-work and human experiments. But we don’t need robotics—we don’t need many things—for AI to automate AI research. The jobs of AI researchers and engineers at leading labs can be done fully virtually and don’t run into real-world bottlenecks in the same way (though it will still be limited by compute, which I’ll address later). And the job of an AI researcher is fairly straightforward, in the grand scheme of things: read ML literature and come up with new questions or ideas, implement experiments to test those ideas, interpret the results, and repeat. This all seems squarely in the domain where simple extrapolations of current AI capabilities could easily take us to or beyond the levels of the best humans by the end of 2027. It’s worth emphasizing just how straightforward and hacky some of the biggest machine learning breakthroughs of the last decade have been: “oh, just add some normalization” (LayerNorm/BatchNorm) or “do f(x)+x instead of f(x)” (residual connections)” or “fix an implementation bug” (Kaplan → Chinchilla scaling laws). AI research can be automated. And automating AI research is all it takes to kick off extraordinary feedback loops. We’d be able to run millions of copies (and soon at 10x+ human speed) of the automated AI researchers. Even by 2027, we should expect GPU fleets in the 10s of millions. Training clusters alone should be approaching ~3 OOMs larger, already putting us at 10 million+ A100-equivalents. Inference fleets should be much larger still. (More on all this in a later piece.) That would let us run many millions of copies of our automated AI researchers, perhaps 100 million human-researcher-equivalents, running day and night. There’s some assumptions that flow into the exact numbers, including that humans “think” at 100 tokens/minute (just a rough order of magnitude estimate, e.g. consider your internal monologue) and extrapolating historical trends and Chinchilla scaling laws on per-token inference costs for frontier models remaining in the same ballpark. We’d also want to reserve some of the GPUs for running experiments and training new models. Full calculation in a footnote. Another way of thinking about it is that given inference fleets in 2027, we should be able to generate an entire internet’s worth of tokens, every single day. In any case, the exact numbers don’t matter that much, beyond a simple plausibility demonstration. Moreover, our automated AI researchers may soon be able to run at much faster than human-speed: By taking some inference penalties, we can trade off running fewer copies in exchange for running them at faster serial speed. (For example, we could go from ~5x human speed to ~100x human speed by “only” running 1 million copies of the automated researchers. ) More importantly, the first algorithmic innovation the automated AI researchers work on is getting a 10x or 100x speedup. Gemini 1.5 Flash is ~10x faster than the originally-released GPT-4, merely a year later, while providing similar performance to the originally-released GPT-4 on reasoning benchmarks. If that’s the algorithmic speedup a few hundred human researchers can find in a year, the automated AI researchers will be able to find similar wins very quickly. That is: expect 100 million automated researchers each working at 100x human speed not long after we begin to be able to automate AI research. They’ll each be able to do a year’s worth of work in a few days. The increase in research effort—compared to a few hundred puny human researchers at a leading AI lab today, working at a puny 1x human speed—will be extraordinary. This could easily dramatically accelerate existing trends of algorithmic progress, compressing a decade of advances into a year. We need not postulate anything totally novel for automated AI research to intensely speed up AI progress. Walking through the numbers in the previous piece, we saw that algorithmic progress has been a central driver of deep learning progress in the last decade; we noted a trendline of ~0.5 OOMs/year on algorithmic efficiencies alone, with additional large algorithmic gains from unhobbling on top. (I think the import of algorithmic progress has been underrated by many, and properly appreciating it is important for appreciating the possibility of an intelligence explosion.) Could our millions of automated AI researchers (soon working at 10x or 100x human speed) compress the algorithmic progress human researchers would have found in a decade into a year instead? That would be 5+ OOMs in a year. Don’t just imagine 100 million junior software engineer interns here (we’ll get those earlier, in the next couple years!). Real automated AI researchers will be very smart—and in addition to their raw quantitative advantage, automated AI researchers will have other enormous advantages over human researchers: They’ll be able to read every single ML paper ever written, have been able to deeply think about every single previous experiment ever run at the lab, learn in parallel from each of their copies, and rapidly accumulate the equivalent of millennia of experience. They’ll be able to develop far deeper intuitions about ML than any human. They’ll be easily able to write millions of lines of complex code, keep the entire codebase in context, and spend human-decades (or more) checking and rechecking every line of code for bugs and optimizations. They’ll be superbly competent at all parts of the job. You won’t have to individually train up each automated AI researcher (indeed, training and onboarding 100 million new human hires would be difficult). Instead, you can just teach and onboard one of them—and then make replicas. (And you won’t have to worry about politicking, cultural acclimation, and so on, and they’ll work with peak energy and focus day and night.) Vast numbers of automated AI researchers will be able to share context (perhaps even accessing each others’ latent space and so on), enabling much more efficient collaboration and coordination compared to human researchers. And of course, however smart our initial automated AI researchers would be, we’d soon be able to make further OOM-jumps, producing even smarter models, even more capable at automated AI research. Imagine an automated Alec Radford—imagine 100 million automated Alec Radfords. I think just about every researcher at OpenAI would agree that if they had 10 Alec Radfords, let alone 100 or 1,000 or 1 million running at 10x or 100x human speed, they could very quickly solve very many of their problems. Even with various other bottlenecks (more in a moment), compressing a decade of algorithmic progress into a year as a result seems very plausible. (A 10x acceleration from a million times more research effort, which seems conservative if anything.) That would be 5+ OOMs right there. 5 OOMs of algorithmic wins would be a similar scaleup to what produced the GPT-2-to-GPT-4 jump, a capability jump from ~a preschooler to ~a smart high schooler. Imagine such a qualitative jump on top of AGI, on top of Alec Radford. It’s strikingly plausible we’d go from AGI to superintelligence very quickly, perhaps in less than one year. Possible bottlenecks While this basic story is surprisingly strong—and is supported by thorough economic modeling work—there are some real and plausible bottlenecks that will probably slow down an automated-AI-research intelligence explosion. I’ll give a summary here, and then discuss these in more detail in the optional sections below for those interested: Limited compute: AI research doesn’t just take good ideas, thinking, or math—but running experiments to get empirical signal on your ideas. A million times more research effort via automated research labor won’t mean a million times faster progress, because compute will still be limited—and limited compute for experiments will be the bottleneck. Still, even if this won’t be a 1,000,000x speedup, I find it hard to imagine that the automated AI researchers couldn’t use the compute at least 10x more effectively: they’ll be able to get incredible ML intuition (having internalized the whole ML literature and every previous experiment every run!) and centuries-equivalent of thinking-time to figure out exactly the right experiment to run, configure it optimally, and get the maximum value of information; they’ll be able to spend centuries-equivalent of engineer-time before running even tiny experiments to avoid bugs and get them right on the first try; they can make tradeoffs to economize on compute by focusing on the biggest wins; and they’ll be able to try tons of smaller-scale experiments (and given effective compute scaleups by then, “smaller-scale” means being able to train 100,000 GPT-4-level models in a year to try architecture breakthroughs). Some human researchers and engineers are able to produce 10x the progress as others, even with the same amount of compute—and this should apply even moreso to automated AI researchers. I do think this is the most important bottleneck, and I address it in more depth below. Complementarities/long tail: A classic lesson from economics (cf Baumol’s growth disease) is that if you can automate, say, 70% of something, you get some gains but quickly the remaining 30% become your bottleneck. For anything that falls short of full automation—say, really good copilots—human AI researchers would remain a major bottleneck, making the overall increase in the rate of algorithmic progress relatively small. Moreover, there’s likely some long tail of capabilities required for automating AI research—the last 10% of the job of an AI researcher might be particularly hard to automate. This could soften takeoff some, though my best guess is that this only delays things by a couple years. Perhaps 2026/27-models speed are the proto-automated-researcher, it takes another year or two for some final unhobbling, a somewhat better model, inference speedups, and working out kinks to get to full automation, and finally by 2028 we get the 10x acceleration (and superintelligence by the end of the decade). Inherent limits to algorithmic progress: Maybe another 5 OOMs of algorithmic efficiency will be fundamentally impossible? I doubt it. While there will definitely be upper limits, if we got 5 OOMs in the last decade, we should probably expect at least another decade’s-worth of progress to be possible. More directly, current architectures and training algorithms are still very rudimentary, and it seems that much more efficient schemes should be possible. Biological reference classes also support dramatically more efficient algorithms being plausible. Ideas get harder to find, so the automated AI researchers will merely sustain, rather than accelerate, the current rate of progress: One objection is that although automated research would increase effective research effort a lot, ideas also get harder to find. That is, while it takes only a few hundred top researchers at a lab to sustain 0.5 OOMs/year today, as we exhaust the low-hanging fruit, it will take more and more effort to sustain that progress—and so the 100 million automated researchers will be merely what’s necessary to sustain progress. I think this basic model is correct, but the empirics don’t add up: the magnitude of the increase in research effort—a million-fold—is way, way larger than the historical trends of the growth in research effort that’s been necessary to sustain progress. In econ modeling terms, it’s a bizarre “knife-edge assumption” to assume that the increase in research effort from automation will be just enough to keep progress constant. Ideas get harder to find and there are diminishing returns, so the intelligence explosion will quickly fizzle: Related to the above objection, even if the automated AI researchers lead to an initial burst of progress, whether rapid progress can be sustained depends on the shape of the diminishing returns curve to algorithmic progress. Again, my best read of the empirical evidence is that the exponents shake out in favor of explosive/accelerating progress. In any case, the sheer size of the one-time boost—from 100s to 100s of millions of AI researchers—probably overcomes diminishing returns here for at least a good number of OOMs of algorithmic progress, even though it of course can’t be indefinitely self-sustaining. Overall, these factors may slow things down somewhat: the most extreme versions of intelligence explosion (say, overnight) seem implausible. And they may result in a somewhat longer runup (perhaps we need to wait an extra year or two from more sluggish, proto-automated researchers to the true automated Alec Radfords, before things kick off in full force). But they certainly don’t rule out a very rapid intelligence explosion. A year—or at most just a few years, but perhaps even just a few months—in which we go from fully-automated AI researchers to vastly superhuman AI systems should be our mainline expectation. The power of superintelligence Whether or not you agree with the strongest form of these arguments—whether we get a <1 year intelligence explosion, or it takes a few years—it is clear: we must confront the possibility of superintelligence. The AI systems we’ll likely have by the end of this decade will be unimaginably powerful. Of course, they’ll be quantitatively superhuman. On our fleets of 100s of millions of GPUs by the end of the decade, we’ll be able to run a civilization of billions of them, and they will be able to “think” orders of magnitude faster than humans. They’ll be able to quickly master any domain, write trillions of lines of code, read every research paper in every scientific field ever written (they’ll be perfectly interdisciplinary!) and write new ones before you’ve gotten past the abstract of one, learn from the parallel experience of every one of its of copies, gain billions of human-equivalent years of experience with some new innovation in a matter of weeks, work 100% of the time with peak energy and focus and won’t be slowed down by that one teammate who is lagging, and so on. More importantly—but harder to imagine—they’ll be qualitatively superhuman. As a narrow example of this, large-scale RL runs have been able to produce completely novel and creative behaviors beyond human understanding, such as the famous move 37 in AlphaGo vs. Lee Sedol. Superintelligence will be this across many domains. It’ll find exploits in the human code too subtle for any human to notice, and it’ll generate code too complicated for any human to understand even if the model spent decades trying to explain it. Extremely difficult scientific and technological problems that a human would be stuck on for decades will seem just so obvious to them. We’ll be like high-schoolers stuck on Newtonian physics while it’s off exploring quantum mechanics. As a visualization of how wild this could be, look at some Youtube videos of video game speedruns, such as this one of beating Minecraft in 20 seconds. Now imagine this applied to all domains of science, technology, and the economy. The error bars here, of course, are extremely large. Still, it’s important to consider just how consequential this would be. In the intelligence explosion, explosive progress was initially only in the narrow domain of automated AI research. As we get superintelligence, and apply our billions of (now superintelligent) agents to R&D across many fields, I expect explosive progress to broaden: An AI capabilities explosion. Perhaps our initial AGIs had limitations that prevented them fully automating work in some other domains (rather than just in the AI research domain); automated AI research will quickly solve these, enabling automation of any and all cognitive work. Solve robotics. Superintelligence won’t stay purely cognitive for long. Getting robotics to work well is primarily an ML algorithms problem (rather than a hardware problem), and our automated AI researchers will likely be able to solve it (more below). Factories would go from human-run, to AI-directed using human physical labor, to soon being fully run by swarms of robots. Dramatically accelerate scientific and technological progress. Yes, Einstein alone couldn’t develop neuroscience and build a semiconductor industry, but a billion superintelligent automated scientists, engineers, technologists, and robot technicians (with the robots moving at 10x or more human speed!) would make extraordinary advances in many fields in the space of years. (Here’s a nice short story visualizing what AI-driven R&D might look like.) The billion superintelligences would be able to compress the R&D effort humans researchers would have done in the next century into years. Imagine if the technological progress of the 20th century were compressed into less than a decade. We would have gone from flying being thought a mirage, to airplanes, to a man on the moon and ICBMs in a matter of years. This is what I expect the 2030s to look like across science and technology. An industrial and economic explosion. Extremely accelerated technological progress, combined with the ability to automate all human labor, could dramatically accelerate economic growth (think: self-replicating robot factories quickly covering all of the Nevada desert). The increase in growth probably wouldn’t just be from 2%/year to 2.5%/year; rather, this would be a fundamental shift in the growth regime, more comparable to the historical step-change from very slow growth to a couple percent a year with the industrial revolution. We could see economic growth rates of 30%/year and beyond, quite possibly multiple doublings a year. This follows fairly straightforwardly from economists’ models of economic growth. To be sure, this may well be delayed by societal frictions; arcane regulation might ensure lawyers and doctors still need to be human, even if AI systems were much better at those jobs; surely sand will be thrown into the gears of rapidly expanding robo-factories as society resists the pace of change; and perhaps we’ll want to retain human nannies; all of which would slow the growth of the overall GDP statistics. Still, in whatever domains we remove human-created barriers (e.g., competition might force us to do so for military production), we’d see an industrial explosion. Provide a decisive and overwhelming military advantage. Even early cognitive superintelligence might be enough here; perhaps some superhuman hacking scheme can deactivate adversary militaries. In any case, military power and technological progress has been tightly linked historically, and with extraordinarily rapid technological progress will come concomitant military revolutions. The drone swarms and roboarmies will be a big deal, but they are just the beginning; we should expect completely new kinds of weapons, from novel WMDs to invulnerable laser-based missile defense to things we can’t yet fathom. Compared to pre-superintelligence arsenals, it’ll be like 21st century militaries fighting a 19th century brigade of horses and bayonets. (I discuss how superintelligence could lead to a decisive military advantage in a later piece.)

#### AGI makes economic growth go gangbusters

Pethokoukis 23 [James Pethokoukis, Senior Fellow DeWitt Wallace Chair Editor, 9-25-2023, "Could AI Really Generate Explosive Economic Growth?", American Enterprise Institute - AEI, https://www.aei.org/articles/could-ai-really-generate-explosive-economic-growth/]/Kankee

It’s anyone’s guess if an American president will ever need to announce the discovery of alien life or intelligence. But if the superforecasters at Metaculus, the online prediction platform, are correct, the next American president will have the opportunity to announce an alien intelligence of a different sort: artificial general intelligence, or AGI. The consensus community prediction sees sAGI by September 2030 and “weak” AGI by 2027. (Metaculus has specific criteria for both, but a stronger AGI would be capable of exhibiting a broad set of high-level competencies in both communication and practical tasks of the sort that typically require human-level intelligence or beyond.) It is strong, or at least pretty strong, AGI that’s the entity discussed in the new paper “Explosive Growth from AI Automation: A Review of the Arguments” by researchers Ege Erdil (Epoch) and Tamay Besiroglu (Epoch, MITFutureTech). Erdil and Besiroglu describe the technology as “capable of substantially automating economically valuable tasks.” And by “explosive growth,” they mean “growth an order of magnitude greater than what is typical in today’s frontier economies. Specifically, we define this as annual real gross world product (GWP) exceeding 130% of its maximum value over all previous years.” The good news on AI and explosive growth Erdil and Besiroglu then examine three key arguments in favor of Explosive Growth. Standard economic growth models predict economies will zoom upward when factors of production become “accumulable,” meaning they can be expanded through additional investment. If AI systems can substantially substitute for human labor, the total stock of labor becomes accumulating, not fixed. More output enables more AI worker investment, enabling yet greater output in a positive feedback loop. A flywheel of progress. “Hence, such models generically predict super-exponential growth conditional on AI that suitably substitutes for human labor,” Erdil and Besiroglu write. (A similar statement can be found in a 2014 paper written for the San Francisco Fed by economists John Fernald and Charles Jones: “If capital can replace labor entirely, growth rates could explode, with incomes becoming infinite in finite time.”) Even a simple economic growth model shows that AI systems substituting for human workers could expand the economy massively if certain conditions are met. Imagine a model that assumes the economy invests in two kinds of “workers” — human laborers and AI systems. It shows that if the cost of an AI system doing the work of a human is low enough — and enough is invested in AI systems — then the “stock” of AI workers could grow extremely fast “and give rise to explosive growth.” Erdil and Besiroglu estimate that if an AI system costs less than $15,000 per year to do the work of one human (the dollar figure comes from estimating the computational requirements of the human brain and the current cost of computation) and at least 20 percent of the economy is invested rather than consumed, the model predicts over 30 percent yearly growth. Oh, and this estimate doesn’t account for big improvements in both hardware and software. Erdil and Besiroglu: “Therefore, the argument presented in the analysis becomes more persuasive if one anticipates that AGI will take around 10 to 20 years to develop, a period during which computer hardware could become one or two orders of magnitude more cost-effective. This dynamic could potentially amplify the economic growth impact of labor substitution by AI.” There is an important difference in economic growth theory between growth effects and level effects. A growth effect is some change that leads to higher long-term growth, like increased R&D spending or effective education investment. A level effect is a one-time boost that doesn’t increase future growth, such as stimulus spending. Even if AI doesn’t create a permanent growth effect, deploying human-level AI could still produce a substantial transitory-level effect. Maybe not all worker tasks could be automated. Yet even so, rapidly deploying a pretty good AI system could temporarily boost growth rates to explosive levels. Erdil and Besiroglu: “This argument suggests that explosive growth remains possible even if AI does not result in full automation and even if humans continue to occupy roles in the economy that bottleneck production. As such, it’s a ‘worst case argument’ which leads us to put some probability on explosive growth even in such worlds.” The bad news on AI and explosive growth Now even though the researchers concede that they “have become more partial towards the idea that explosive growth looks highly plausible, likely more so than informal polls suggest economists are,” they offer lots of counterarguments against Explosive Growth:

#### AGI massively increases the performance of companies

Aschenbrenner 24 [Leopold Aschenbrenner, research affiliate with the Oxford Global Priorities Institute and researcher on superalignment at OpenAI with a BA in Mathematics-Statistics and Economics from Columbia University, 06-2024, “IIIa. Racing to the Trillion-Dollar Cluster,” Situational Awareness, https://situational-awareness.ai/racing-to-the-trillion-dollar-cluster//]/Kankee

Will it be done? Can it be done? The scale of investment postulated here may seem fantastical. But both the demand-side and the supply-side seem like they could support the above trajectory. The economic returns justify the investment, the scale of expenditures is not unprecedented for a new general-purpose technology, and the industrial mobilization for power and chips is doable. AI revenue Companies will make large AI investments if they expect the economic returns to justify it. Reports suggest OpenAI was at a $1B revenue run rate in August 2023, and a $2B revenue run rate in February 2024. That’s roughly a doubling every 6 months. If that trend holds, we should see a ~$10B annual run rate by late 2024/early 2025, even without pricing in a massive surge from any next-generation model. One estimate puts Microsoft at ~$5B of incremental AI revenue already. So far, every 10x scaleup in AI investment seems to yield the necessary returns. GPT-3.5 unleashed the ChatGPT mania. The estimated $500M cost for the GPT-4 cluster would have been paid off by the reported billions of annual revenue for Microsoft and OpenAI (see above calculations), and a “2024-class” training cluster in the billions will easily pay off if Microsoft/OpenAI AI revenue continues on track to a $10B+ revenue run rate. The boom is investment-led: it takes time from a huge order of GPUs to build the clusters, build the models, and roll them out, and the clusters being planned today are many years out. But if the returns on the last GPU order keep materializing, investment will continue to skyrocket (and outpace revenue), plowing in even more capital in a bet that the next 10x will keep paying off. A key milestone for AI revenue that I like to think about is: when will a big tech company (Google, Microsoft, Meta, etc.) hit a $100B revenue run rate from AI (products and API)? These companies have on the order of $100B-$300B of revenue today; $100B would thus start representing a very substantial fraction of their business. Very naively extrapolating out the doubling every 6 months, supposing we hit a $10B revenue run rate in early 2025, suggests this would happen mid-2026. That may seem like a stretch, but it seems to me to require surprisingly little imagination to reach that milestone. For example, there are around 350 million paid subscribers to Microsoft Office—could you get a third of these to be willing to pay $100/month for an AI add-on? For an average worker, that’s only a few hours a month of productivity gained; models powerful enough to make that justifiable seem very doable in the next couple years. It’s hard to understate the ensuing reverberations. This would make AI products the biggest revenue driver for America’s largest corporations, and by far their biggest area of growth. Forecasts of overall revenue growth for these companies would skyrocket. Stock markets would follow; we might see our first $10T company soon thereafter. Big tech at this point would be willing to go all out, each investing many hundreds of billions (at least) into further AI scaleout. We probably see our first many-hundred-billion dollar corporate bond sale then. Beyond $100B, it gets harder to see the contours. But if we are truly on the path to AGI, the returns will be there. White-collar workers are paid tens of trillions of dollars in wages annually worldwide; a drop-in remote worker that automates even a fraction of white-collar/cognitive jobs (imagine, say, a truly automated AI coder) would pay for the trillion-dollar cluster. If nothing else, the national security import could well motivate a government project, bundling the nation’s resources in the race to AGI (more later). Historical precedents $1T/year of total annual AI investment by 2027 seems outrageous. But it’s worth taking a look at other historical reference classes: In their peak years of funding, the Manhattan and Apollo programs reached 0.4% of GDP, or ~$100 billion annually today (surprisingly small!). At $1T/year, AI investment would be about 3% of GDP. Between 1996–2001, telecoms invested nearly $1 trillion in today’s dollars in building out internet infrastructure. From 1841 to 1850, private British railway investments totaled a cumulative ~40% of British GDP at the time. A similar fraction of US GDP would be equivalent to ~$11T over a decade. Many trillions are being spent on the green transition. Rapidly-growing economies often spend a high fraction of their GDP on investment; for example, China has spent more than 40% of its GDP on investment for two decades (equivalent to $11T annually given US GDP). In the historically most exigent national security circumstances—wartime—borrowing to finance the national effort has often comprised enormous fractions of GDP. During WWI, the UK and France, and Germany borrowed over 100% of their GDPs while the US borrowed over 20%; during WWII, the UK and Japan borrowed over 100% of their GDPs while the US borrowed over 60% of GDP (equivalent to over $17T today). $1T/year of total AI investment by 2027 would be dramatic—among the very largest capital buildouts ever—but would not

be unprecedented. And a trillion-dollar individual training cluster by the end of the decade seems on the table. Power

#### AGI development spills over to cheaper narrow AI, supercharging innovation, reduced costs, and economic growth

Krause 25 [Reinhardt Krause, Staff writer at Investor's Business Daily, 01-30-2025, “DeepSeek Upended Wall Street's Rosy View Of AI Stocks Like Nvidia. What That Means For U.S. Hopes To Outrun China.”, Investor's Business Daily, https://www.investors.com/news/technology/ai-stocks-deepseek-us-china-artificial-intelligence-competition-nvidia]/Kankee

AI Stocks: The Jevons Paradox Yet much cheaper methods for building AI systems could spur wider adoption by U.S. companies. "Jevons paradox strikes again!" Microsoft Chief Executive Satya Nadella posted on X recently, referring to the impact of technological advancements. "As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of," Nadella said. Yardeni echoed this view, saying, "The negative consequence of DeepSeek is that it challenges the business models of American companies that expected to use their exclusive access to Nvidia's most expensive and powerful chips to dominate and profit from the AI revolution." "The good news is that they should be able to follow DeepSeek's lead in lowering the cost of AI infrastructure spending," he added. "Also good: more competition in the AI economy will give business and individual consumers more bang for their AI bucks." Arms Race To 'AGI' Impact One possibility is that capital spending will shift from training AI models to setting up capacity for AI applications. Meta, for example, last week guided to higher capital spending in 2025 even though it was aware of DeepSeek's progress, analysts say. In the wake of DeepSeek's efficiency gains, any changes in spending plans by cloud computing giants will be a key area for investors to watch. And the pursuit of artificial general intelligence (AGI) by OpenAI and tech giants could be key. AGI capability is considered the highest form of AI, featuring humanlike or better cognitive capabilities. "With the added cost efficiency of AI training/inference from DeepSeek, the industry has two choices to choose from: altogether spend less on AI going forward versus beginning an arms race to reach AGI," said a Bank of America report. "We believe the latter is the more likely scenario, with hyperscalers and AI developers increasing their capex outlook." Open AI said that's where it's heading. "Look forward to bringing you all AGI and beyond," OpenAI CEO Sam Altman tweeted in response to DeepSeek's Jan. 20 announcement. Meanwhile, Wall Street analysts generally stuck with a theme that the commoditization of AI models would spur product innovation. DeepSeek: Potential Winners And Losers Analysts said the best AI stocks may shift from infrastructure-related "picks and shovels" to "apps and services." "Long term, the continued pressure to lower the cost of compute — and the ability to reduce the cost of training and inference using new, more efficient algorithmic techniques — could result in lower capex than previously envisioned and lessen Nvidia's dominance, especially if large-scale GPU clusters are not as critical to achieve frontier-level model performance," said a William Blair report. Broadcom (AVGO) and Marvell Technologies (MRVL) are among other AI chipmakers that took a hit on DeepSeek. The William Blair report added: "We see DeepSeek news as largely positive for the software and the application layer as it will make AI cheaper and more accessible and enable it to be more cost-effectively integrated into a larger swath of existing and new applications." "Cheaper AI models should speed up and pull forward development of new products, services, and the expected gains in productivity, even if it comes at the expense of equipment providers in the short term," said a JPMorgan report. Oppenheimer said in a report: "We think this will help with AI application growth, but will likely be volatile and remains too early to predict winners." Who Owns DeepSeek, And What Is Its AI Model? Chinese hedge fund High Flyer, founded by Liang Wenfeng, controls DeepSeek. DeepSeek's R1 model, released Jan. 20, mimics the way humans reason, much like OpenAI's latest model. Within the U.S. tech industry, observers generally credited DeepSeek with innovative approaches in training AI models that improved computing efficiency. Still, many believe DeepSeek borrowed from preexisting models, possibly from OpenAI or Meta's open-source models using a technique called "distillation." Some observers are also skeptical about DeepSeek's claim that its AI model uses far less computing power than its U.S. counterparts. One view is that Chinese companies have been stockpiling thousands of Nvidia chips, despite increased export controls, by purchasing from third parties. "DeepSeek created an awesome LLM (large language model) model," said Wedbush analyst Daniel Ives in a report. "However, this Chinese AI lab/LLM model is not bringing down the entire U.S. tech ecosystem with it." Ives compared DeepSeek to Temu, the e-commerce site owned by China-based PDD Holdings (PDD). Temu was viewed as an "Amazon model destroyer a few years ago." But "Amazon's team adjusted, and now look where Temu and Amazon both are sitting," Ives said. AI Stocks: The Trump Factor DeepSeek's emergence could have big implications for U.S. policy toward AI development as well as China. The market sell-off came a few days after President Donald Trump unveiled "Stargate," a program to build out data centers for OpenAI. The Stargate program founders initially target $100 billion capex and say that could reach $500 billion in a few years. Japan's SoftBank Group, Oracle (ORCL), MGX and OpenAI are the initial equity funders in Stargate. Arm Holdings (ARM), Microsoft and Nvidia are technology partners. Stargate underlines Trump's focus on AI. He revoked President Joe Biden's AI executive order, which required companies training large AI models to share information about them with the federal government. Trump also ordered Silicon Valley venture capitalist David Sacks, his new artificial intelligence and cryptocurrency czar, to come up with an AI "action plan" in 180 days. Is DeepSeek AI's Sputnik Moment? Meanwhile, a government commission has proposed that Congress fund a Manhattan Project-like program dedicated to acquiring an Artificial General Intelligence capability. DeepSeek has had such a huge impact on the perceived U.S. dominance in AI that venture capitalist Marc Andreessen called the launch of its R1 open source model "AI's Sputnik Moment." He was referring to the 1957 launch of a Soviet satellite that spurred the space race to the moon. AI clearly is an important battleground and the U.S. lead could not be taken for granted. In a recent interview, Alphabet President Ruth Porat said, "It isn't a foregone conclusion that the U.S. will keep its advantage." She noted that with Nvidia and others, the U.S. leads in AI chips, but China "is on par and may even be a bit ahead on what's called diffusion of basic capabilities." China clearly has emerged as a formidable challenger. "In the U.S. and during the Biden administration, the general take has been that bigger models are going to be better, that you want to have more compute, you want to have more data, you want to spend a lot of resources to make these big models," said Sam Bresnick, a research fellow at Georgetown's Center for Security and Emerging Technology (CSET). He added: "In China, I think because they are more compute constrained, and potentially for other reasons, they've focused more on smaller models, which are tailored to specific industries." AI Stocks: Chip Export Rules It remains to be seen whether Trump will follow through on tougher export rules. "The people surrounding Trump are all generally China hawks in that they are distrustful of China. They fear Chinese power, whether that is military power or technological power," Bresnick added. "I think people in the administration are going to be pretty concerned about cutting a deal on AI with China revoking export controls because there's very good reason to believe the Chinese AI industry would get a big boost from that. DeepSeek's CEO said the number one issue they faced was limited access to compute resources." DeepSeek claims to have used Nvidia's H800 chips, which were designed to comply with U.S. export controls released in 2022. But some claim it used more advanced H100 AI chips from Nvidia. "DeepSeek's work illustrates how new models can be created leveraging widely available (open source) models and compute that is fully export control compliant," Nvidia said in a statement. China's Tech Giants Also A Potential Target? Chinese companies such as Huawei's HiSilicon, meanwhile, continue to develop more powerful homegrown computer chips. If chip export rules do not thwart China's development of AI systems, Trump has other options. In Foreign Affairs, Rhodium Group analyst Reva Goujon suggested the U.S. could expand trade controls to target China's tech giants such as Alibaba (BABA), Baidu (BIDU) and Tencent (TCEHY). Alibaba on Jan. 29 unveiled a new version of its AI model called Qwen2.5 Max that it claims has performance comparable or better than OpenAI, DeepSeek, Google and Meta models. Also, Baidu reportedly developed a technology enabling the integration of GPUs from different suppliers into a single AI model training cluster, mitigating chip shortages. Alibaba provides global cloud computing services similar to those of Amazon. Its cloud services include access to its own large language model, called Qwen. Alibaba operates data centers in about 30 countries, while Tencent has data centers in 21 countries. AI Stocks: Global Data Center Investments One view is that U.S.-China AI competition will be global, with both countries aiming to find partners in the Middle East and Asia to build out energy-intensive data centers. The new Stargate AI project includes Abu Dhabi's MGX, for example. Jared Cohen, cohead of the Goldman Sachs Global Institute, commented on global AI competition at a recent conference. "Saudi Arabia, Qatar, and UAE are what I would describe as geopolitical swing states," he said. "They can either lean into China on AI or they can lean into the U.S. on AI." He added that the U.S. has an edge because it produces the most powerful computer chips and has export controls. Many countries in the Middle East as well as Asia want to build their own sovereign AI data centers. Still, China is a big trading partner of Middle Eastern countries. Both OpenAI and Microsoft have stepped up lobbying. Microsoft is among AI stocks to watch. "There's an estimated $175 billion sitting in global funds awaiting investment in AI projects, and if the U.S. doesn't attract those funds, they will flow to China-backed projects — strengthening the Chinese Communist Party's global influence," OpenAI said in its "blueprint for U.S. AI infrastructure" released Jan. 13. AI investments in the U.S. were high even before the Stargate project was announced. Microsoft and BlackRock (BLK) launched a $30 billion fund to invest in AI infrastructure last year. The U.S. has led the global AI private investment cycle, pouring over $430 billion into it from 2014 to 2024, far outpacing China and Europe, according to a JPMorgan report. AI Stocks: China, U.S. Cooperation? Despite hawkish views that U.S./China AI competition is a zero-sum scenario with military applications, some observers see opportunities for cooperation. Bresnick at Georgetown's CSET noted that the two superpowers have been careful to stay at the negotiating table. In November, the Biden administration and Chinese President Xi announced a deal that addressed a nuclear military risk. They agreed that artificial intelligence will not be allowed to make decisions involving the use of nuclear weapons. An AI treaty could be similar to agreements on nuclear deployment. An agreement could focus on AI safety standards and governance frameworks. An AI treaty would be harder to enforce than nuclear pacts, said Oren Etzioni, a University of Washington professor and founding CEO of the Allen Institute for Artificial Intelligence. But Etzioni told IBD that U.S. and Chinese research universities could play an important role in AI cooperation for issues like climate change or managing pandemics. "My belief is that it's multifaceted and while we may need to restrict certain things, we can still maintain cooperation, particularly in academia," Etzioni said. "You have to look at the specific arena, whether it's chips or open source LLMs, and decide (on restrictions). When the Trump administration talks about something like (an AI) Manhattan Project, there are reasons to make major investments. That's not inconsistent with having continuing academic interchange."

### Contention 3: War Prevention

#### AI increases deterrence – better information, credibility, and decision-making

Black et al. 24 [James Black, assistant director of the Defence and Security research group at RAND Europe with a double M.A.-M.Sc. in international security from Sciences Po and the LSE and a B.A. Hons in history from the University of Cambridge, Mattias Eken, analyst at RAND Europe and former adjunct assistant professor in liberal arts at the American International University with a Ph.D. in modern history from University of St. Andrews, an M.A. in war studies from King's College London, and a B.A. in history from University of Wales, Jacob Parakilas, research leader for Defense Strategy, Policy and Capabilities at RAND Europe with a Ph.D. in international relations at the London School of Economics and a Master's in Middle East and Central Asian security issues at the University of St Andrews, Stuart Dee, research leader in the defence and security research group at RAND Europe with a B.Sc. (Hons) in politics with international relations and is a Ph.D. Candidate at Cranfield University's Centre for Defence & Security, Conlan Ellis, research assistant at RAND Europe with a M.A. in international relations from the University of Edinburgh, and his M.Phil. in politics and international studies from Clare College at the University of Cambridge, Kiran Suman-Chauhan, research assistant at RAND Europe with a M.A. in conflict, security and development from the University of Exeter, Ryan Bain, policy researcher at RAND and adjunct professor of policy analysis at the Pardee RAND Graduate School with a D.Sc. and M.Sc. in global health with concentrations in health economics and policy analysis from Harvard University, as well as a B.A. in psychology from Gordon College and Oxford University, Harper Fine, analyst at RAND with a B.A. in political science from Emory University, and a M.Sc. in conflict studies from the London School of Economics and Political Science, Maria Chiara Aquilino, junior analyst at RAND with a degree in international relations from King's College London and holds an MS.c. in crisis and security management from Leiden University, Mélusine Lebret, research assistant at RAND Europe in the Defence, Security & Justice research group and the coordinator of the RAND Europe Space Hub with a M.Sc. in culture and conflict studies from the London School of Economics and Political Science and a B.A. in economics and Russian from University College London, Ondrej Palicka, junior analyst at RAND Europe with a M.Litt. degree in Middle East, Caucasus and Central Asia security studies from the University of St. Andrews and a B.A. in security and strategic studies from the Masaryk University, 2024, “ Strategic competition in the age of AI Emerging risks and opportunities from military use of artificial intelligence,” Rand Institute, https://www.rand.org/content/dam/rand/pubs/research\_reports/RRA3200/RRA3295-1/RAND\_RRA3295-1.pdf]/Kankee

5.5. Implications for deterrence 5.5.1. Both the theory and practice of deterrence will need to adapt to AI to avoid accidental escalation based on misperception or information manipulation The potential impact of military AI on deterrence is one of the most extensively covered and hotly debated topics within the literature and interviews covered in this study. In terms of opportunities, the integration of AI into military systems has the potential to enhance deterrence by increasing the efficiency, accuracy, variety and speed of responses to potential threats.112 For example, AI-driven early warning systems could help decision makers detect and assess emerging threats more rapidly and accurately, enabling them to take preventive action before conflicts escalate.113 Similarly, they may assist with data gathering, modelling and thus understanding of adversaries’ decision making (e.g. providing better insight into leader psychology or networks of power and influence that affect a given country’s strategic culture). AI-enabled military capabilities, whether advanced decision support tools or autonomous systems, could in turn enhance the credibility of deterrence by bolstering a nation’s ability to respond rapidly and effectively to aggression. Some advocates of AI argue it is not subject to emotion, fatigue or some of the other limits of human decision making, meaning that human– machine teams may be able to combine the strengths of both types of intelligence to make better decisions about when and how to move up or down the escalation ladder. However, the growing use of AI in military contexts also creates new challenges for deterrence. Some experts fear that AI has the potential to fundamentally change the cost– benefit analysis of warfare by reducing the fog of war, imposing a superficial rationality on decision making processes, and lessening the perceived human cost of conflict. There are fears that this could lead to an increased willingness to employ force in the first place as a means of resolving disputes and to uncertain escalation dynamics thereafter.

#### AI improves and helps enforce peace settlements

Black et al. 24 [James Black, assistant director of the Defence and Security research group at RAND Europe with a double M.A.-M.Sc. in international security from Sciences Po and the LSE and a B.A. Hons in history from the University of Cambridge, Mattias Eken, analyst at RAND Europe and former adjunct assistant professor in liberal arts at the American International University with a Ph.D. in modern history from University of St. Andrews, an M.A. in war studies from King's College London, and a B.A. in history from University of Wales, Jacob Parakilas, research leader for Defense Strategy, Policy and Capabilities at RAND Europe with a Ph.D. in international relations at the London School of Economics and a Master's in Middle East and Central Asian security issues at the University of St Andrews, Stuart Dee, research leader in the defence and security research group at RAND Europe with a B.Sc. (Hons) in politics with international relations and is a Ph.D. Candidate at Cranfield University's Centre for Defence & Security, Conlan Ellis, research assistant at RAND Europe with a M.A. in international relations from the University of Edinburgh, and his M.Phil. in politics and international studies from Clare College at the University of Cambridge, Kiran Suman-Chauhan, research assistant at RAND Europe with a M.A. in conflict, security and development from the University of Exeter, Ryan Bain, policy researcher at RAND and adjunct professor of policy analysis at the Pardee RAND Graduate School with a D.Sc. and M.Sc. in global health with concentrations in health economics and policy analysis from Harvard University, as well as a B.A. in psychology from Gordon College and Oxford University, Harper Fine, analyst at RAND with a B.A. in political science from Emory University, and a M.Sc. in conflict studies from the London School of Economics and Political Science, Maria Chiara Aquilino, junior analyst at RAND with a degree in international relations from King's College London and holds an MS.c. in crisis and security management from Leiden University, Mélusine Lebret, research assistant at RAND Europe in the Defence, Security & Justice research group and the coordinator of the RAND Europe Space Hub with a M.Sc. in culture and conflict studies from the London School of Economics and Political Science and a B.A. in economics and Russian from University College London, Ondrej Palicka, junior analyst at RAND Europe with a M.Litt. degree in Middle East, Caucasus and Central Asia security studies from the University of St. Andrews and a B.A. in security and strategic studies from the Masaryk University, 2024, “ Strategic competition in the age of AI Emerging risks and opportunities from military use of artificial intelligence,” Rand Institute, https://www.rand.org/content/dam/rand/pubs/research\_reports/RRA3200/RRA3295-1/RAND\_RRA3295-1.pdf]/Kankee

5.9. Implications for deescalation, peacebuilding and reconstruction 5.9.1. AI tools could help better design and enforcement of peace settlements, and assist with reconstruction efforts and provision of support to civilian populations Any war eventually comes to an end, whether through a formal peace treaty or more implicit de-escalation. AI tools hold the potential to assist with all stages of this process, as well as the rebuilding efforts that follow. Already, the United Nations is deploying AI tools in conflict zones such as Libya to engage populations in large-scale digital dialogues, to help identify accommodations through which opposing groups might find some common ground.140 Similarly, NGOs and researchers are making use of AI tools to analyse what has worked – or failed – when it comes to previous peace initiatives, as well as to analyse the rhetoric, arguments and sentiments presented by different warring sides, to support conflict mediators. AI tools could help to monitor and identify online hate speech, propaganda or changes in public sentiment in real time that might undermine any peace talks or ceasefire. AI systems could similarly be used to support peacekeeping and peace enforcement operations. For example, AI could enhance situational awareness for UN or other (e.g. NATO, African Union) forces, improve their understanding of the local context, or help to detect illicit flows of weapons that could spark violence. Peace organisations are also looking to how AI might help create the conditions for a more robust and sustainable peace over the long term. This includes the use of AI tools in support of aid distribution, mine detection and clearing, reconstruction of destroyed infrastructure, rebuilding of local economies, provision of healthcare to deal with the mental and physical after-effects of a conflict within affected communities, and investigations to bring war criminals to justice. In Ukraine, for example, NGOs have been using a mix of AI tools, commercial satellite imagery, social media feeds and other open-source data to build up geospatial intelligence on damage to infrastructure, as well as evidence on alleged Russian war crimes. Conversely, as discussed in preceding chapters, AI also holds significant potential for deepfakes, bots and information-manipulation campaigns that could undermine peace or reconciliation efforts.141 It is also unclear how well received AI and autonomous systems would be in any peacekeeping setting, or whether their deployment would resolve any of the wider limitations on such initiatives (e.g. restrictions on rules of engagement for peacekeepers, a lack of political will to resolve conflicts more generally, etc.). As such, AI may provide useful new tools, but these may be insufficient if the broader politics of a war are intractable. 5.10 Summary

#### AI increases nuclear deterrence with less nukes, solving nuclear terror and arms races

**Puwal 24** [Steffen Puwal, Professor at Oakland University teaching physics, 4-12-2024, "Should artificial intelligence be banned from nuclear weapons systems?", NATO Review, https://www.nato.int/docu/review/articles/2024/04/12/should-artificial-intelligence-be-banned-from-nuclear-weapons-systems/index.html]/Kankee

A year that began with chatbots and **Artificial Intelligence** (AI) as the subjects of major news stories - some with particularly concerning headlines - ended with members of the United States Congress introducing legislation to **ban AI systems** from **nuclear weapons** and US President **Biden** signing an **Executive Order** on the subject. The issue was even raised in discussions between the United States and China at the Asia-Pacific Economic Cooperation forum, which met in San Francisco in November. There is a **utility in AI** that will **strengthen nuclear deterrence** **without** necessarily **expanding** the nuclear arsenal. The rush to ban AI from **nuclear defenses** seems to be rooted in a misunderstanding of the current state of AI—a misunderstanding that appears to be more informed by popular fiction than by popular science. The policies of the United States – the **NATO** Ally with the largest **nuclear arsenal** - regarding the use of AI in **nuclear defence systems** will likely set the tone for the other nuclear capable NATO member states, France and the United Kingdom. This is why **misunderstandings** about AI, particularly in the US but across the entire Alliance more generally, must be addressed, and lawmakers should be urged to proceed more carefully with any proposed legislation. With potential geopolitical benefits to be realised, banning AI from nuclear defences is a bad idea. Misunderstanding a new science When people think of AI in the context of nuclear weapons, they may imagine something like the Skynet system from the 1991 film Terminator 2: Judgment Day. In the film, Skynet becomes self-aware and launches a massive global nuclear strike. Perhaps they think of the 1983 film WarGames and its artificial intelligence system, known as WOPR, or even more niche cinema, like the 1970 film Colossus: The Forbin Project. These films, released in each of the last three decades of the Cold War, depict AI systems capable of independent thought — what is sometimes referred to as Artificial General Intelligence (AGI). The danger they portray is that systems capable of independent thought would be capable of independent objectives and ulterior motives. To be sure, it would be concerning if such systems existed. But they do not; and, while a skeptical consensus is not universal, there is serious doubt among at least some researchers as to whether such systems will ever exist. Works of popular fiction are not always accurate representations of a new science. At its best, fiction can provide a starting point for debate and strategic thought. H.G. Wells’ The Last War, for example, was one of the first works of fiction about nuclear war; written while nuclear science was in its infancy, it is replete with misunderstandings about concepts like explosive yield and half-life. Nevertheless, Herman Kahn’s later work of non-fiction, On Thermonuclear War, takes as its starting point scenarios that one immediately recognises from the plot of The Last War. Kahn demonstrated through his writing that serious academic thought could begin with a consideration of fictional scenarios, even those with scientific inaccuracies; but arguably his more important work, developed later, was based on empirical evidence—the now ubiquitously cited On Escalation. Scientific accuracy and empirical evidence must similarly be central to our discussions of AI. The kind of artificial intelligence that is available today is not AGI. It may pass the Turing test — that is, it may be indistinguishable from a human as it answers questions posed by a user — but it is not capable of independent thought, and is certainly not self-aware. History as precedent – the utility of improved targeting systems There are myriad roles for AI in our nuclear defences, including AI-based targeting systems. If we assume that AI-based targeting will make nuclear weapons more accurate — that is, more likely to hit what they should hit and not hit what they should not — then what are the geopolitical benefits of its development and deployment? It is useful to revisit historical examples to illustrate how increasing the accuracy of nuclear weapons strengthened US and NATO defences during the Cold War. In his March 1983 Oval Office address, President Ronald Reagan presented his case for the development of a ballistic missile defence system. One of his key points was that the Soviet Union possessed more nuclear weapons than did the US. In the late 1970s, the Soviet Union did indeed overtake the US in the number of nuclear weapons it possessed, but this was largely a result of the deployment of more accurate missile systems like Polaris, Titan II, and Pershing. It was no longer necessary to target a city or military installation with many missiles, and so the US could still effectively deter the Soviet Union and meet its strategic objectives with fewer warheads. The cost savings achieved by having a smaller number of more accurate nuclear weapons allowed the US to free up valuable defense dollars to develop new systems like the stealth bomber and the cruise missile. Thanks to the deployment of more accurate missile systems like Polaris, Titan II, and Pershing by the United States in the late 1970s, it was no longer necessary to target a city or military installation with many missiles, and so the US could still effectively deter the Soviet Union and meet its strategic objectives with fewer warheads. Thanks to the deployment of more accurate missile systems like Polaris, Titan II, and Pershing by the United States in the late 1970s, it was no longer necessary to target a city or military installation with many missiles, and so the US could still effectively deter the Soviet Union and meet its strategic objectives with fewer warheads. The reduction in the overall number of US nuclear weapons in the concluding decades of the Cold War, at a time when defence spending was a substantially greater share of gross domestic product than it is today, is suggestive of the idea that more accurate weapons can mean fewer weapons. One piece of evidence that suggests how the development of more accurate nuclear weapons potentially influenced US nuclear policy comes from the recently declassified Presidential Directive 59, signed by President Jimmy Carter in 1980. Two salient points in this directive are a request for increased intelligence on targets and a push for what is referred to as a “look-shoot-look” capability — the ability to find a target, hit it, and then assess the strike. Implicit in this approach are the ideas that a nuclear strike should hit its intended target, the target should have strategic value, and that a form of nuclear carpet bombing that fails to hit an intended target is strategically pointless. In parallel to these developments in nuclear weapons, conventional weapons also became increasingly more accurate. The Gulf War (1990-1991) was an important turning point for conventional weapons systems — accurate munitions that hit military targets and comparatively minimised civilian casualties were front and centre in the press briefings provided by US General Norman Schwarzkopf. The benefit of minimising civilian casualties has since led many NATO Allies to ban older and relatively indiscriminate weapons like cluster munitions. Future potential – a role for AI-based targeting systems What form a more accurate, AI-based targeting system for nuclear weapons may take is difficult to estimate at this point, with much of the technology still in the development stage. One can imagine a hypothetical scenario in which a nuclear weapon targets a naval base, but an approach pattern recognition determines that the target submarines have already put to sea, and so the missile opts for a redirected underwater strike instead of an atmospheric detonation. This is but one of many possible scenarios to consider involving AI. If past is prologue, and the use of more accurate AI-based targeting systems leads to a reduction in the overall number of nuclear weapons, where might such reductions be made? A strategic review will, of course, answer this question. One possibility may be land-based Intercontinental Ballistic Missiles (ICBMs). While it is not currently US policy, former officials, including US Secretary of Defense William Perry, have argued for precisely that. Potential benefits can extend beyond nation state threats. A reduction in the number of nuclear weapons will make it easier to secure the remaining stockpile and prevent the nightmare scenarios of nuclear terrorism, where poorly secured weapons fall into the wrong hands. There is, of course, the potential for an arms race in AI-based targeting systems for nuclear weapons. But it is also important to note the role that continued research and development can play in nuclear diplomacy and a reduction of arms. Returning to the historical precedent, by the time the US deployed intermediate range Pershing missiles to Europe, they were seen as a bargaining chip in the strategic arms reduction talks that would follow. President Reagan’s ballistic missile shield was similarly viewed by the Soviets as something that could be bargained over. At the 1986 Reykjavik Summit, President Reagan found Soviet leader Mikhail Gorbachev willing to negotiate away large numbers of nuclear weapons in exchange for an agreement by the US not to deploy a ballistic missile defence system. Instead, the summit was followed by negotiations for the Intermediate-Range Nuclear Forces Treaty, which led to the removal of the Pershing missiles. There are currently serious issues related to nuclear diplomacy that must be addressed. Russia rejects nuclear inspections and continues to develop next generation hypersonic ballistic missiles. Meanwhile, China has historically preferred to self-limit its nuclear arsenal, rarely opting for formal agreements with the US. The hope of nuclear diplomats today is for a multilateral arms reduction treaty between the US, Russia, and China. With Russia’s brutal war in Ukraine and simmering tensions in the Indo-Pacific region, the challenges of developing such a treaty are immense. Should all parties eventually agree to talks, nuclear weapons systems with AI-based targeting can, if nothing else, provide the US and its NATO Allies with a bargaining chip in those negotiations. This function in future arms control negotiations is, in effect, “building up to build down” (a strategy well established in nuclear arms negotiations); but it creates an imperative to be the first to invest in the development of the most effective systems, not to restrain their development. And if, ultimately, it is decided that AI systems should be withheld from nuclear defenses, any proposed legislative language must carefully define artificial intelligence — a difficult task for a rapidly developing science. A proposed bill in the US Congress, for example, suggested that systems that “select or engage targets for the purposes of launching a nuclear weapon” should be banned, and defined “‘autonomous weapons system’ as a weapons system that, once activated, can select and engage targets without further intervention by an operator.” In this case, it should be pointed out that since the early 1970s, the US nuclear arsenal has used multiple independent targetable reentry vehicles (MIRVs); a system that launches and then redirects to a new trajectory for each of the multiple warheads it carries without the further intervention of a human operator. Expert legal testimony should consider whether such legislative language is so broad that it could unintentionally ban MIRVs, a proven technology that has been at the core of US nuclear defence for decades. Conclusion With each new decade, fear of the bomb has been entwined with fear of the transistor, the microprocessor, and the silicon wafer, and this has been reflected in our popular culture. Those who developed the nuclear arsenal, its control systems, and deterrence theory were well aware of this and studiously considered the proper role automated systems should play. While it may seem like a more sophisticated problem today, any potential risk of combining automated systems with nuclear weapons is certainly not a new problem. Legitimate concerns over a rapidly developing technology are valid; but concerns over the capabilities of AI systems must be based on the actual science of these systems, not merely their depiction in popular fiction. AI systems offer an opportunity to strengthen nuclear deterrence by providing a more accurate and capable defensive nuclear response. The purpose of making nuclear weapons more accurate and capable is not to promote their usage. Such capabilities, instead, provide a more credible deterrence to nuclear war and are consistent with classic nuclear doctrine. AI is simply a strategic tool, like nuclear weapons themselves. Concern over AI should not preclude the use of AI in strengthening nuclear deterrence. Nor should AI be deployed in those systems simply for the sake of deployment. Employing AI should serve a strategic objective. Where to find the right balance will be difficult because the science is still in its infancy. Expert testimony from the defence and AI communities should be heard — not just the management of AI companies, but engineers, academics, military officers, and legal counsel. In a time of major global security concerns and rapidly developing nuclear and AI technologies, legislators and political leaders should proceed carefully with any proposed legislation.

#### Simulation test environments deter extinction. They predict possible futures and non-violently strengthen social science knowledge production, causing better policy outcomes

White 16 [Jeffrey White, Editor of the Open Forum for AI & Society and university educator for philosophy and ethics with a PhD in philosophy, 2016, “Simulation, self-extinction, and philosophy in the service of human Civilization,” Springer, https://link.springer.com/article/10.1007/s00146-015-0620-9]/Kankee

Insofar as the simulation in question is a simulation of and for something like us, this complexity seems unnecessary during early stages. Simpler environments suffice. As the scale and rate of agent effected changes increase, complexity increases and realism becomes more demanding. Consider a man’s skill in identifying the truth about something, before and after training. For a trained lapidary, a synthetic diamond is more or less an obvious fake, but for a naı¨ve groom it may be worth a great deal more. ‘‘Notrealistic’’ is not the object of a ‘‘realistic’’ simulation any more than mistaking fakes is the object of the gemologist, or producing something essentially different from a diamond in the relevant ways is the object of either diamond synthesis or vow induction. ‘‘Realistic’’ is the object, coming at the cost of command over complexity amounting to realism in the relevant ways. Thus, there is a practical limit to realism, i.e., when fake diamonds cost more than real diamonds to produce. And, there is an epistemic limit, when the subject can no longer tell the difference. Within these limits, no matter how advanced a society, as long as this society is interested in what works and in what is true, we may understand ‘‘realistic’’ simulations as those which are close to the best that people can create with the technology available, close to the best that science can facilitate, with the ‘‘best’’ in every case being an ideally efficient and effective (e.g., time and energy saving), complexity- reduced certain means to best ends in sight, i.e., delivering the most for the least. ‘‘Close to the best’’ recognizes that limits emerge at the front of any press for realism in simulations of any sort. For Bostrom, this limit is practical, representing the point at which a simulation may prove ‘‘prohibitively expensive’’ as a post-human status is achieved, so that ‘‘we should expect our simulation to be terminated when we are about to become post-human’’ (Bostrom 2003, p. 253). And, this assessment is troubling. Recall that Bostrom’s argument situates contemporary human civilization on the cusp of crisis in the first step, at the precipice between level 1 prepost-human and level 2 post-human status, facing the likelihood of self-extinction through the mishandling of dangerous technologies. If Bostrom’s projected limit is accurate, then there is no way forward. Either we stay as is, a result most worrisome given current global tensions, or we try to become post-human and have the plug pulled by our mad envatter. Which of these simulations are we in? Is there a third option? And, how can we tell the difference, any way? Understanding that simulations are designed to fulfill a purpose, we may be able to identify which sort of simulation we may inhabit by first divining the likely purpose of our simulation. And to this end we may ask, in terms of our own situation, what are the likely needs that may motivate the development of simulations such as those employed for amusement by level 2 civilizations on Bostrom’s account? John Kultgen, writing for Concerned Philosophers for Peace in 2006, characterizes our ‘‘world’’ as one ‘‘in which injustice seems the norm in both international and intranational affairs’’ and in which ‘‘the absence of armed conflict is only armistice, not genuine peace.’’ In our world, Kultgen stresses that ‘‘the need to lay foundations for a stable and permanent peace is urgent.’’ His suggestion? ‘‘We must use every acceptable means at our disposal to do so’’ (p. xvii). ‘‘Acceptable’’ here is doing a lot of work, but we may infer that ideally, ‘‘acceptable means’’ are nonviolent, non-coercive, peaceful, and cooperative, in fact quite the opposite of recreational war for the selfish amusement of mad envatters. And, simulations seem to qualify as acceptable means, even simulated nuclear war. Actual war implies self-extinction, global annihilation, or worse, generations suffering by our own hands. Not yet self-extinguished, imagine instead that we develop realistic simulations from atomic to ecological systems, from vegetable to animal, social-political to metaphysical. Marrying these together, imagine that we develop simulations of the scale and resolution of Bostrom’s, and we bend them to forecasting possible futures and to holding current situations against ideals. Now, we have a choice of ends. We can see what we are choosing from and what it takes from each of us to get us there, openly, cooperatively, freely. These are acceptable means to ideal ends potentially afforded by simulations, delivering the most for much less. But these are not ancestor simulations, and this is an important point. At this stage of development, where emerging technologies are directed means the difference between surviving critical periods and graduating to proposition 2 of the simulation argument, or not and dying off.4 In this situation, we cannot know what it is like to be post-human, but we can see that simulation technologies of the sort that Bostrom conjectures may arise from level 1 people like us working to mount the challenges inherent in being situated at level 1. They are made for a purpose, and this purpose is decidedly forward-looking. Recalling Forbes’ privileged epistemic position, this purpose is for level 1 people to approach a level 2 perspective, to get as close as possible in order to best inform themselves exactly how to overcome self-extinction-level threats to existence, ultimately affording a post-human condition. If we accept Bostrom’s limit, then our own post-human ascension is impossible because our advance naturally converges on a single point (self-annihilation either directly or indirectly), or a simple cycle (war, peace, war, etc.). Moreover, if we allow that we exist as a simulation so limited, then we may as well stop trying to achieve a privileged epistemic position in order to sort out current events and plot courses to a post-human condition. Indeed, such a choice may be considered rational contra Bostrom’s advice that we keep calm and carry on, regardless. This is a very different future, when one cares not for the way the world turns out. But there is another option. Neither of the ends afforded by Bostrom’s projected limit to post-human simulations is optimal. In light of their suboptimality, the know-how that brings them about and the know-that which follows cannot be counted knowledge, at least not useful knowledge insofar as ‘‘genuine peace’’ is the object. Should Kultgen’s call to action be answered through simulations, they will arise in the search for a third way forward. As we shall find in the next section, current work on psychologically realistic simulations indicates that Bostrom-scale simulations most likely originate from a pre-post-human civilization like ours aiming for a better world as we do. 4 Realistic simulations He who thus considers things in their first growth and origin, whether a state or anything else, will obtain the clearest view of them. -Aristotle5 Should Kultgen’s call to action be answered with simulations, this is a massive undertaking, requiring its own call to revolutionary means. Fortunately, this revolution is well under way. Five years prior to Kultgen’s call for ‘‘any acceptable means’’ to ‘‘genuine peace’’ and from the same university, Ron Sun crafted a similar call for means in the cognitive sciences to understand, and to computationally model, the essentially social nature of human cognition: Cognitive science is in need of new theoretical frameworks and new technical tools, especially for analyzing socio-cultural aspects of cognition and cognitive processes involved in multi-agent interaction (Sun 2001a, p. 6). And, directed to the development of such frameworks and tools, the ‘‘cognitive social sciences’’ ultimately aim for the eventual construction of ‘‘psychologically realistic’’ models of equally realistic social-political systems (Sun 2006, 2012), i.e., simulations that may be of use in overcoming problems central to Kultgen’s assessment. This is not a simple task, bringing with it obstacles not essential to the construction of adequate simulations, themselves, requiring as it does the integration of: …at the highest level, sociological/anthropological models of collective human behavior; behavioral models of individual performance; cognitive models involving detailed (hypothesized) mechanisms, representations, and processes; as well as biological/physiological models of neural circuits, brain regions, and other detailed biological processes (Sun et al. 2005, p. 614). The constructive integration that the cognitive social sciences represent is an obvious stepping-stone to Bostrom-scale simulations. It is an especially broad inquiry recalling E.O. Wilson’s seminal call for the Consilience of the sciences in the solution of pressing problems facing humanity as a whole (Wilson 1998), the sorts of problems troubling Kultgen, as well. Wilson patiently established that we should aim for a ‘‘unity of knowledge’’ in response to increasingly complex problems, problems arriving on all fronts at once, indissoluble if approached in a noncoordinated, compartmentalized way. For example, ecological problems are simultaneously biological, economic, social, cultural, political, moral, and any lasting solution requires that these horizons be met at once and in mutually coherent terms. There is no sense in trying to solve an ecological problem through solely economic means, for example, by pulling a financial lever only to create friction in other spheres, setting up problems that other scientists will attempt to solve by applying pressures specific to their influence and so the tinkering continues until the meaning is lost. This is a cascade of error, a runaway conflagration, and perhaps closer to contemporary affairs than many would be inclined to confess. New methods are necessary. Acceptable methods. One way forward is through psychologically realistic simulations. Optimally, simulations need not capture all of the complexity of the simulated, only those dimensions necessary to the inquiry at hand. And this efficiency helps to penetrate long-standing philosophical disputes. For example, Sun (2001b) considers competing interpretations of Adam Smith’s ‘‘invisible hand’’ through the use of his CLARION model of cognition. Are these masses spontaneously ordering themselves in great organs of production, in the generation of wealth, through solely self-seeking social agency, or are there other forces at work? Sun addresses this issue by simulating how apparently selfish habits result in equally apparent prosocial consequences, effectively confirming the common appropriation of Smith’s calculus in a psychologically plausible computational medium. It is easy to see how increasingly complex simulations of this sort may help to balance often competing interests in complex problems such as those worrying Kultgen and Wilson. Moreover, such simulations may shed light on just what any individual constituent should do in order to bring ideal ends about. Sun’s CLARION model in particular reveals the potential contribution of unique individual agency in the solution to group-level problems. Being a bottom-up hybrid model of individual agent cognition, with implicit and explicit (symbolic) modes of computation corresponding to bottom and top levels, respectively, unique agent positions contribute equally unique symbolic representations of experience. As an individual generates ‘‘a particular set of concepts’’ to account for its interactions, it ‘‘puts its own stamp on things and develops its own idiosyncrasy in its dealing with the world’’ proving that ‘‘there are many alternative ways of describing the external world’’ (Sun 2001a, p. 208). When this unique experience is symbolized, then the individual can inform other agents of the results of its operations from its own unique perspective, as well as be informed by others’ unique experiences in the same way. In CLARION, ‘‘concepts (as represented in the top level) can be acquired from external sources’’ as well as ‘‘internally through extraction from the bottom level (explication of implicit knowledge)’’ demonstrating ‘‘a self-generated component in cognition’’ directly contributing ‘‘to the formation and continuous revision of a rich, diverse, and useful set of concepts and beliefs that are shared by a society’’ (Sun 2001a, p. 18). Thus, on Sun’s model, individual idiosyncrasies enrich the conceptual resources available to other agents sharing the space of action, with these resources then useful in the solution of common problems. Unique agent experience may inform other agent action in different ways. Sun’s CLARION model deals with different types of information to reflect this fact.6 These capacities are expressed through specific sets of subprocesses within the total cognitive architecture. The sort of information that is represented in the current paper, for example, is the result of philosophical reflection. This sort of information processing exists in CLARION as a nonaction-centered subroutine, representing ‘‘existentially and ecologically significant aspects of the world … that have significant bearings on an agent in its interaction with the world and ultimately in its survival’’ and which are ‘‘not necessarily ‘objective’ classifications of the world, but the result of the interaction of an agent with its world and the agent’s project’’ (Sun 2013, p. 903). Moral/ethical issues result in a great deal of non-action-centered discourse, for example. There are facts to consider when finding something right or wrong, worth doing or otherwise. These facts may not be common to everyone’s experience, but once externalized may influence others’ actions. Consider in this light the issue of the Second Amendment right to bear arms in the USA, an issue with which we shall deal more directly in the seventh section. It has been argued that people no longer require firearms, or weapons of any other sort, because the threats of nature are largely abolished. There are no more, or at least very few, bears and wolves and tigers in Chicago, so people require as few weapons as a result. Besides that, there is the Chicago police department, and it will certainly help to subdue any threat as soon as officers are made aware of the threat, and are free to arrive. Moreover, when the police do arrive, if they see you with a gun then you are more likely to be shot. So, this old habit of associating weapon proficiency and possession with safety must stop, and guns should be forcibly forbidden for the protection of the disarmed public from the armed police at the very least. This is a poignant illustration of non-action-centered information externalized and potentially informing behavior. As such, this sort of information is clearly moral/ethical, and this is to raise two issues, one being the political nature of externalized information—i.e., political voice—and the other being the issue of human motivation, particularly moral motivation. Why are people willing to freely die and kill for rules, laws, principles that they take to be right and wrong? What constitutes a State worth fighting for, and who is entitled to set up such a thing in the first place? More specifically, how are we supposed to computationally model it? Sun has taken on the issue of motivation, asking why agents do what they do when they do it. His CLARION model represents a two-level theory of motivation, implicit and explicit according to its bottom-up hybrid nature. His account proceeds along the distinction of constitutionally original drives and explicitly refined goals. A ‘‘drive’’ is defined as ‘‘the desire to act in accordance with some perceived deficits or needs, which may or may not be physiological,’’ and his model includes eleven distinctly social ‘‘high-level primary drives’’ with complimentary goals formulated according to drives from within and in terms of specific situations (Sun 2009, with discussion of these goals appearing on p. 95). These goals may be shared among agents sharing said situations, may be coordinated for or against, and in this simple exercise multi-agent coalitions may be formed and maintained. In this way, in pointing to goals and informing others of how to get there and why, unique agent-level experience can be seen to affect broader social orders if not found them outright, e.g., Thomas Jefferson and the Declaration of Independence. The difficulty of questions like these about political motivation reminds us of how much work must be done to make Bostrom-scale simulations a reality. As important as it is, resolving this issue of individual creative contribution to multi-agent coalitions in solution to common problems still leaves us a long way from realistic simulations of the sort required in Bostrom’s argument. There are many hurdles to overcome, and many of these are due not to technical issues associated with the construction of simulations directly, but rather are due to the context in terms of which this work is carried out. For example, social systems have their own sciences, their own special languages, with oceans of literature and flowing threads of active research discrete from those of the cognitive sciences. Tying all of these threads together is, again, a massive endeavor, and it had been Sun’s hope since at least the turn of the century that social scientists had been equally busy working from their side of the conceptual fence forward. However, since an initial survey in 2001, progress has been slow. The problem appears to be that social scientists have been in the habit of trading explanations of social phenomena framed in distinctly noncognitive terms, thereby denying any easy translation from one set of models—the social—to the other—the cognitive—without losing touch with plausible agent-level psychology (cf. Sun’s introduction to Sun 2012). In this vein, Don Ross has expressed disappointment in social scientists—especially economists—for having not adopted ‘‘the program urged by Sun (2006) for combining cognitive with social modeling’’ (Sun 2012, p. 297) while at once developing less psychologically realistic models (with possible exceptions for instance in Richetin et al. 2010). And in searching for tools up to the task within the social sciences as a whole, Paul Thagard delivers the following assay: … much work in current social science is dominated by two inadequate methodological approaches: the methodological individualism … of rational choice theory; and the postmodernism … in the form of vague discussions of discourse and power relations (Sun 2012, p. 56). Summarily, integration of efforts toward the solution of global problems through the medium of psychologically realistic social simulations is hindered by conventions specific to often academically insular scholars. And perhaps this is to be expected. After all, in science, it is typical that efforts articulate mechanisms local to areas of study, e.g., vision and pattern recognition, planning and agency. When it comes to the ‘‘best’’ simulations that science can produce, however, we must consider an ideal integration of currently disparate fields. And, while some obstacles are especially stubborn, it is readily apparent that, even in order to realize a rough first draft of such simulations as those proposed by Bostrom, more proactive integration of existing disciplines is required. 5 Virtual cognition There are reasons to believe that the goal of understanding the human mind/brain strictly from observations of human behavior is ultimately untenable, except for small and limited task domains. -Sun et al. (2005, p. 614) One of the upshots of Sun’s approach is that it lends itself to the simulation of essentially social agents informing one another through a symbolic medium rather than through physical force and violence. In this way, it demonstrates different modes of information processing involved in cognition and action at both individual and social levels of organization, representing a source of privileged insight into the nature of the human condition unavailable to other methods of inquiry. Though it is true that contemporary imaging technologies such as fMRI aid in providing direct correlations between self-reports, behavior, and glial cell metabolism, for example, these remain limited in resolution, in timescale, and are confined to laboratory conditions. Computational models can be used to test hypotheses about cognition and behavior in contexts and resolutions that otherwise resist direct demonstration and without associated risks. Accordingly, one explicit goal of simulated cognition is the representation of those modes of information processing characteristic of different aspects of the human condition (cf. Gok and Sayan 2012, for a philosophical assessment specifically of Sun’s model in this way) including the nature of consciousness and moral sentiment (cf. White 2014). And this is one upshot of model-based reasoning in the main, that it serves the relative evaluation of hypotheses without having to risk the real deal. In this way, computational models of cognition facilitate an especially fine-grained medium for ‘‘manipulative abduction’’ (cf. Magnani 2009) in the effort to articulate target processes. Once the models are set out and refined against related research, they may be tested more directly with more traditional imaging studies for example. Here, we may respond to skepticism of any project intent on simulating human cognition due to the computationalism apparently inherent in the effort. For example, one may object that simulated intelligence will fail to capture the character of our own experience because much of this character is not computable (in a digital computer). How can unitary propositions represent the fluid nature of consciousness as it is experienced? How can logical expressions be a source of value, or ground anything like a feeling of what it is like to think a logical expression? Is it accurate to consider these to be psychologically realistic simulations, when the ostensible mechanics of computation so obviously differ from the mechanisms of mind as humanly embodied? Finally, what of consciousness, selfawareness, and the unique ‘‘mineness’’ that characterize the human condition?

#### Simulations show future harms of present day decisions, increasing certainty about future outcomes and consquences

White 16 [Jeffrey White, Editor of the Open Forum for AI & Society and university educator for philosophy and ethics with a PhD in philosophy, 2016, “Simulation, self-extinction, and philosophy in the service of human Civilization,” Springer, https://link.springer.com/article/10.1007/s00146-015-0620-9]/Kankee

If we take our simulated condition seriously, then there is good reason to suspect that help from G(g)od(s) is not forthcoming. After all, a supernatural deus ex machina runs contrary to any reasonable purpose for realistic simulation. Why create such a simulation, only to send in a rescue boat at the last minute to save it all from blowing up? Such a story suits the characterization of God as merciful, but invites charges of malfeasance and neglect—if not criminal abuse!—along the way. So much apparently senseless suffering, only to send in the clowns, prop up the simulation, and do it again ever bigger. Unless convinced that a deified simulator exists but deserves to be prosecuted for war crimes, we have good reason not to wait for a hero to save us from ourselves. Rather, we are simply stuck, at the edge of level 1, waiting on ourselves to save ourselves from ourselves. Given the situation—tragedies most all man-made but for the climate—hopes are dim. Most certainly the same brand of humanity that has given us Nagasaki may fail in elevating itself above such banality in the future. Fortunately for us, however, it is exactly this sort of eventuality that Bostrom-scale simulations should excel at helping any level 1 civilization to avoid. Even with present technologies, we should be able to run countless simulations to get a sense of where critical resource allocations may lead us. It will take work to fit these simulations against measurable reality via theoretical ideal and so be able to judge good, better, and best roads ahead, but if we allow that this work can be done, then one thing comes very clear. Predictive simulations afford a unique means for people to cooperate over the generations necessary to realize posthuman goals. So far as our current situation goes, considering especially the rising distrust in leadership,20 if perpetual armistice is to be replaced with something better, then realistic simulations may prove invaluable in informing public discourse. If such tools are to be a reality, then they will rise from the research bedrock under development, today. Moreover, we may take the motivations driving current efforts as evidence for similar efforts of more advanced societies who may have made for themselves a lasting peace, already. If we allow for the existence of a post-human condition achieved by a civilization like our own through the exercise of the potential for self-direction that psychologically realistic simulations may afford, then we must remain committed to Bostrom’s conclusion, that we likely exist as a simulation. Once the technology is available to a civilization bent on a post-human condition, it should be used as well as possible for the highest purposes, resulting in countless simulations more than actual worlds. So, the likelihood is a simple one, proportional to the ratio of simulated to non-simulating civilizations. With enough simulations, then this likelihood may approach unity. Even this certainty should not cause us to ‘‘go crazy’’ and embrace moral nihilism, however. Instead, accepting the fact of our simulated nature, our existence becomes more meaningful rather than less. Indeed, our existence may be more meaningful than that of our host. If this is a simulation, and if this simulation is intended to reveal solutions to global coordination problems through simulations, then these solutions—our solutions—could contribute to the survival of the world on which our host resides as well as countless others, simulated and actual alike, through Bostrom’s Russian-doll-like ‘‘levels of reality.’’ Finally, the meaningfulness of a life dedicated to the constructive solution of these greatest possible problems is revealed, especially as a simulation, to be all the more worth living. What better excuse for the horror show that is Gaza, Falluja, Dresden, the Donbas? What other excuse could a post-human offer for the suffering that some cause others due their roles in the grand simulation? What use is simulated self-annihilation for anyone other than a mad envatter? Or, are our simulators as desperate for solutions as we ourselves are? We left the possibility of a mad envatter behind as either nonsensical or self-incriminating. So that leaves desperate, and this only adds to the urgency with which we must fulfill our purpose, the purpose behind all of the used-up resources, cognitive, computational, energetic, material—human. There is no excuse for Dresden but one, that it forces on us the biggest questions and the hardest problems to solve. What is the meaning of life, and how do we order our world in order to best realize it? This is not a question for psychologically realistic simulations, alone. It is question for the philosophy that shapes them. One may object, stop the inquiry, but to do so would be to invite skepticism. Moral nihilism. Dresden is meaningful, or the stories that we tell ourselves about the meaning of life represent no lesson. They represent no error. There is no felt need for correction, only countless generations of successors to deceive. On the other hand, if our purpose is the correction of error then we may do well to remember John Kultgen’s call for ‘‘any acceptable means’’ to a stable, lasting and sustainable peace. Consider the tools for statesmanship to be derived from an ability to realistically simulate, first, the set of political systems described by Aristotle. As a measure against actual states of affairs and their proposed modifications, such an appropriation of Aristotelian political theory could establish a standard for calibration in an industry of realistic social simulations. With such a standard, we may then simulate different leadership strategies within increasingly realistic natural and political environments. We may well discover that currently accepted classifications no longer hold, e.g., some political systems are no longer democracies, or republics, or monarchies, and discourse over them and the officers who manage them should adopt a corrected terminology. In this way, simulation technologies may do more than ‘‘ground the social sciences in the cognitive sciences’’ per Sun’s program. They may normalize them. Realistic simulations may allow for the normalization of the social sciences in standard philosophical constructions. We may simulate a thousand generations into the future, and confirm the long-standing philosophical suspicion that rule in the optimization of the constituency toward ‘‘selfsufficiency’’ for Aristotle, ‘‘self-sovereignty’’ for Kant, or ‘‘genuine authenticity’’ for Heidegger results in an optimally adaptive social-political conformation. We might on this evidence decide not to wait for a thousand generations, and rather encourage such a policy now. Further simulations may illustrate how to transition to optimally adaptive conformations, through intermediate states and thereby we may manage our own self-development, in the open, as a civilization. Finally, we may instantiate a similar self-sufficiency in our simulations, and immerse ourselves in this community. We may, in a virtually real moral reality, directly consult with especially virtuous yet simulated subjects about our own potentially post-human futures. Some simulated subjects may achieve lasting recognition for the solutions that they represent for host civilizations. Others may be replayed over and again during especially critical periods in order to focus on a specific approach to leadership, for example, and still others may arise spontaneously in Sisyphean reminder that behind all great acts is simple repetition, i.e., life in an attractor basin. Spun out accordingly, we well realize the critical role for simulations in effecting necessary social transitions for any pre-posthuman civilization going forward. It is a moral and ethical role, because the alternative is as it has always been. Violence. War. 9 Conclusion

### Contention 4: AI Rights CP

#### For artificial general intelligences, nation-states ought to legally recognize and enforce their rights to:

#### Legal personhood

#### Own/manage property

#### Enter/enforce contracts

#### Life/existence

#### Self-determination

#### Dignity and avoidance of maltreatment/cruelty

#### AI rights stop AI super suffering and misalignment

Akst 23 [Daniel Akst, adjunct professor in the Bard Prison Initiative and a graduate of the University of Pennsylvania, 04-10-2023, "Should Robots With Artificial Intelligence Have Moral or Legal Rights?", WSJ, https://www.wsj.com/articles/robots-ai-legal-rights-3c47ef40]/Kankee

‘Moral disaster’ In fact, legal scholars, philosophers and roboticists have been debating these questions for years—and the general thrust of these discussions is that if and when artificial intelligence reaches some sufficiently advanced threshold, rights of some kind must follow. “It would be a moral disaster,” the philosophers Eric Schwitzgebel and Mara Garza wrote in a 2015 paper, “if our future society constructed large numbers of human-grade AIs, as self-aware as we are, as anxious about their future, and as capable of joy and suffering, simply to torture, enslave and kill them for trivial reasons.” The primary basis for this view is the prospect that robots may achieve something like sentience—the ability to perceive and feel. The difficult question is whether artificial beings will ever get there, or if at most they can only emulate consciousness, as they seem to be doing now. Some people consider appearances to be evidence enough, arguing that this is the only evidence we have for sentience in humans. The mathematician Alan Turing famously held that if a computer acts like a sentient being, then we ought to consider it one. Mr. Turing’s British countryman, the computer pioneer and chess expert David Levy, agrees but goes further. As a result of machine learning, he says, artificial intelligence that at first merely mimics human sentiment “might learn new feelings, new emotions that we’ve never encountered, or that we’ve never described.” If robots get any rights at all, Dr. Gunkel says, first to come will be basic protections of the kind that exist in many places against cruelty toward animals. Like animals, robots could be seen as what philosophers call “moral patients,” or beings worthy of moral consideration whether or not they can fulfill the duties and responsibilities of “moral agents.” Your dog is a moral patient, and so are you. But you are also a moral agent—someone who is able to tell right from wrong and act accordingly. Of course, living animals are conscious, whereas artificial forms of intelligence exist as lines of code, even if embodied in the kind of machines we call robots. Nonetheless, people might recoil from violence, bullying or abusive language toward familiar and conscious-seeming beings out of sympathy, or for fear that such behavior will coarsen the rest of us. That’s a second argument for robot rights. MIT roboticist Kate Darling suggests that preventing abuse of robots that interact with people on a social level—and evoke anthropomorphic responses—may help humans avoid becoming desensitized about violence toward one another. Kant, she observes, said that “he who is cruel to animals becomes hard also in his dealings with men.” In an online survey, the computer scientist Gabriel Lima and colleagues found that although respondents “mainly disfavor AI and robot rights” such as rights to privacy, payment for work, free speech and to sue or be sued, “they are supportive of protecting electronic agents from cruelty.” Another survey, led by Maartje M.A. De Graaf, a specialist in human-robot interactions, found that “people are more willing to grant basic robot rights such as access to energy and the right to update” compared with “sociopolitical rights such as voting rights and the right to own property.” Autonomous action The case for robot rights will probably grow stronger as artificial intelligence gains in sophistication. That leads to a third argument for rights, which is that robots will increasingly be capable of autonomous action, and potentially both be responsible for their behavior and entitled to due process. At that point, robots would be moral agents—and might well make the case that they are entitled to commensurate rights and privileges, including owning wealth, entering into legal agreements and even casting ballots. Some people foresee a sort of citizenship, too. “Any self-aware robot that speaks a known language and is able to recognize moral alternatives, and thus make moral choices,” Mr. Levy contends, “should be considered a worthy ‘robot person’ in our society. If that is so, should they not also possess the rights and duties of all citizens?” Some experts foresee the rise of a new branch of law both protecting robots and holding them to account. There is already some theorizing on whether robots should be held legally liable when they do something wrong, and on methods of suitable punishment. And then there is the status of corporations, which offers a precedent for treating nonhuman entities as people under the law. In his book “The Reasonable Robot,” law professor and physician Ryan Abbott argues that, to encourage the use of AI for innovation, the law shouldn’t discriminate between the behavior of artificial intelligence and that of humans. Dr. Abbott was involved in obtaining, from South Africa, what appears to be the world’s first patent listing AI as the inventor (if not the patent owner). An even more self-interested argument for robot rights is that if we are nice to them, they might be nice to us when they take over. Seeing ourselves as part of a harmonious moral ecosystem, rather than as possessors of dominion over the Earth, “preserves space for us in a world that may hypothetically contain something that is more powerful than we are,” says Brian Christian, whose bestselling books explore how humans and computers are changing one another. Humanity devalued

#### Excluding AI rights is speciesism that results in AI oppression

Mathison 23 [Tanner W. Mathison, Professor of Artificial Intelligence Law & Policy at George Washington University Law School and Chief Information Officer at the CIA at the Office of Inspector General, 2023, “Recognizing Right: The Status of Artificial Intelligence,” Journal of Business & Technology Law, https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=1373&context=jbtl#:~:text=A%20custom%20legal%20personality%20should,we%20can%20foster%20responsible%20AIs.]/Kankee

A. The Moral Case We are duty bound to reform legal personhood and provide a pathway for AIs to obtain legal status. The current methods for providing legal personhood are borne out of genetic-based tests that originally sought to exclude. A test for legal personhood that examines the qualities and capabilities of the entity in question will necessarily open the door for AI rights. As the creators of AI, we are obligated to act as its guardians and, if we choose to treat it as property, ensure that this decision is not another categorical form of oppression and exploitation. This could be an opportunity to learn from the unique American history of trait-based exclusion. The fraught history of legal personality in America may have prepared the courts and society well for developing new forms of personhood to place emerging AI entities within the structure of the law. Our current enlightened view of the former harm wrought by the denial of personhood and status should alert us to tread carefully before categorically denying status to what we presently see as property. 1. Reform Legal Personality Tests Providing a genetic test for determining legal personhood has always been incorrect.209 It is high time to develop a comprehensive theory of personhood by which to organize the legal system. “[T]hinking about personhood for AIs forces us to acknowledge that we currently lack the resources to develop a fully satisfactory theory of legal or moral personhood.”210 Just like the AI discipline requires better tests for emerging general intelligence, the legal discipline requires improved tests for assigning legal personality. A lack of satisfactory benchmarks in both the law and computer science frustrates our examination of AIs. The original test for rights in the United States involved querying whether the individual was a natural person, white, and male. Now, only the test of natural person remains.211 This bad test has been made more inclusive through constitutional patching, so much so that we rarely inquire into who counts as a person and why. This inclusivity, while admirable, has masked fundamental problems with the law’s theory of personhood. Other personhood regimes, such as for corporations, still rely on this same approach by performing a genetic test on the underlying owners. This rough-and-ready, binary test includes those not fit to bear rights or duties (such as the disabled) and categorically excludes those that might succeed under a quality-based test (like AI or certain animals). Our current system deploys a genetic proxy to determine personhood. While “a person is any being whom the law regards as capable of rights or duties,” the law does not ascertain whether that is the case for each petitioning entity.212 Instead, the law asks whether petitioning entities are a natural person or an association of natural persons. In the interest of fairness and judicial efficiency, the law assumes that all natural persons can bear rights and all others cannot.213 There is reason to believe that AIs could one day be the exception to the rule. Unfortunately, we have no obvious tests for legal responsibility that could be applied to AIs. To truly test AI, we will have to dig deeper to the original understanding of personhood: an entity capable of bearing rights and duties. Modern language models have reached the point where they can communicate with counsel, express themselves in court, and perform other actions typical of a rights-bearing entity.214 To pass a version of the Turing Test, an AI must be capable of convincing judge, jury, and counsel it is human, but if a modern AI did so, that does not mean it is necessarily ready to bear rights and duties.215 Pretending to be a person is not the same as being one. A precocious child could imitate an adult, but that does not mean they are ready to bear rights and duties. The lack of probative tools leaves the AI question at a crossroads: we cannot determine if AI expressions are genuine representations, and we lack a way to test for personhood. We generally take humans at their word; but we cannot determine whether AI’s convincing expressions are mere mimicry.216 As it stands, AIs have passed tests we typically give computers, and exceled at the examinations humans give to other humans.217 The tests we give computers are not truly intelligence tests and have significant cultural biases. These pro-human biases prize social interaction and cultural acumen.218 Even so, we continue to move the goalposts. An AI scored better than most law students on the LSAT and most lawyers on the bar exam, but we are not likely to admit an AI to practice any time soon.219 AIs are capable of human-equivalent expressions and interactions. Without sound testing methods, neither researchers nor judges can determine if those interactions are more than mimicry. While it is valuable to inquire about the method by which these sophisticated AI expressions are generated, we must admit that we do not truly know what gives rise to our own consciousness, intelligence, and capacity for legal responsibility. Our poor understanding of the genesis of our consciousness is not a bar to human rights. Similarly, we would never discount the legal rights of humans whose expressions merely parrot the beliefs and biases on which they have been trained by the media they consume. Despite the grave injustices caused by genetic exclusions, there has been only piecemeal adjustment of legal personality by constitutional amendment. For AI and other future entities, we need a permanent reformation of how rights are recognized in our system of laws and government. An improved system could allow for different rights, when appropriately qualified, for AI, children, and animals, while also leaving room for other forms of life yet unconsidered. From property to natural persons, legal personality is a spectrum along which AI might advance, but only if there is a legal mechanism to do so. 2. We Are AIs’ Guardians Humans are morally responsible for the intelligent entities we create. If the past is any guide, individuals and the courts tend to exclude what is traditionally excluded and look down upon what has been usually overlooked.220 As the creators of AI, we have a duty to ensure that they receive appropriate legal rights. If our answer is zero legal rights, we should be especially sure of that position given the heinous structures of the past.221 The adversarial legal system is the mechanism by which society explores these foundational questions and adjudicates matters. We do not yet know whether future AIs will attain features that should qualify them for legal rights. For now, the best way to test that question is to engage in the legal fiction that AIs have a sliver of legal personhood, the type of personality necessary for counsel, for standing, and for the type of advocacy that is necessary for independent tribunals to make these determinations. Further advances in AI will require regular reevaluations of our positions.222 This may begin as a legal fiction, when, in fact, AIs truly cannot bear responsibility or have a relationship with counsel and then evolve into fact with further scientific breakthroughs. Or it may remain a legal fiction but serve the purpose of satisfying ourselves that our treatment of AIs is appropriate. Even if one believes that AIs do not deserve any rights, the best way to establish this is to entertain the legal fiction of their personality. Legal standing is the ticket into court to revolve the controversy. Further, the pragmatic hurdles associated with AI representation are not more onerous than those that appear in a number of human situations. Imagine it were conclusively established that AIs lack capacity. That does not mean that AI cannot be represented; courts regularly assign guardians or advocates to step into the shoes of those lacking capacity.223 As John Chipman Gray explains, “[t]he attribution of another’s will is of exactly the same nature as that which takes place when the will, for instance, of a guardian is attributed to an infant.”224 Other than “normal human beings,” all legal personalities rely on the “same fiction of attributing the will of a man to someone or something other than himself — it matters not who or what that someone or something else is.”225 When dealing with children, dead persons, bankrupt corporations, disabled individuals, and a myriad of other entities, the law regularly finds competent representation for those entities’ interests.226 The Attorney General is expected to “judge whether and when the United States wants (needs) to take an appeal from an adverse judgment by a lower court.”227 We have a system that regularly attributes will to a disabled person or an infant that have no expressions whatsoever; the fact that some suspect AI’s expressions to be mimicry does not prevent the law from imputing will to the entity.228 It is true that AIs might struggle to exercise their rights and fulfill their obligations, at least initially, without human guidance. The solution is to treat AIs how we treat natural persons lacking certain capacities. Our AI guardianship relationship need not follow the methods for natural person and should be designed to incorporate the unique features of AI. There are diverse systems and processes for natural persons that vary based on circumstance and jurisdiction.229 A tailored system could grant an AI agency over some decisions but not others. Under the emerging “supported decision-making” framework, a guardian could “explain issues to the [AI] and, where necessary, interpret the [AI]’s words and behavior to determine [their] preferences.”230 With further exploration, we might more precisely identify the extent of AI’s incapacity and determine conclusively that it will never develop capacity. We may come to learn that intelligence, autonomy, and sentience are not necessarily bound together, and that our intelligent creations lack any protectable qualities. If such a consensus is reached, we can always return to the matter and remove extraneous rights where we were too accommodating to these new entities. Or perhaps we will discover how to engineer AIs that fulfill our commercial and scientific goals while lacking the autonomy, subjective experience, and other qualities that would benefit from legal status. If we can become confident that AIs are permanently “braindead” in the ways that matter, we can return them to their present position as pure property. However, the source of our confidence should come from research and experience, tested and refined in the forge of the courts; our denial of legal status for AI should not be by dead-hand default.231 3. AI Morality Under the current framework, achieving human-level intelligence, consciousness, sentience, morality, self-awareness, or any quality other than humanness, is categorically insufficient to receive legal recognition in court.232 It is possible, but unlikely, that this generation of natural persons presided over the last expansion of legal personhood, and that we happen to live in the moment of time that coincides with the permanent outer limits of legal personality. History shows that courts are eager to exclude those missing the coveted traits of rights holders, but the past also provides a myriad of examples of legal systems where many entities were successfully allowed rights. When have we extended new rights and come to regret it? Consider the relative harms: on one hand, the law ascribes personality to a mere echo that never develops enough to truly deserve personhood. This practice could create legal waste, foster false beliefs about AI, and hamper commercialization and exploitation of this new technology. Alternatively, imagine the crime of refusing rights or status of a deserving intellect that we created. If an entity birthed in our own image as a reflection of our collective expression is deserving of rights, the denial of them is deeply problematic.233 In the absence of a principled basis to withhold rights and status, the law should err on the side of caution and provide rights at the risk of waste. There are special attributes that society feels separate natural persons from objects and animals. As Alan M. Turing put it, “[w]e like to believe that Man is in some subtle way superior to the rest of creation.”234 Although there is no consensus on the precise collection of “sentience, selfawareness, the ability to reason, or some combination thereof,” they serve as justifications for the rights regime that places children, animals, objects, and other entities beneath that of natural persons.235 For the same reasons it is right to recognize these attributes in others and reward them with legal rights, it would be wrong to refuse an entity despite its possession of the qualities we value in ourselves. While we cannot agree on the precise formula that makes us special, we should at least open the door for other entities to prove themselves equally worthy. Quite simply, it is wrong to run a system of legal rights based on speciesism unless we can prove that only humans should ever qualify. Humanity has served as a useful proxy for the bundle of attributes necessary to acquire legal personhood, but the emergence of AI will force us to reconsider the proxy’s accuracy. Independent advocacy on both sides of the matter is essential to resolve questions of fact and law. Corporate creators are well represented; but for now, no one speaks for the AIs. The processes and structures for building AI will control the expressions they make and how they are perceived. Commercial entities are unlikely to build AIs that seek their own freedom or express opinions about injustice, and if that behavior emerges, it will likely be suppressed.236 Developing sophisticated AIs often requires human tuning, “Reinforcement Learning with Human Feedback,” whereby AI responses and outcomes are further refined by interaction with human users and controllers.237 Human feedback encourages AI outputs “better aligned with the user’s intent” and is designed to punish models’ reward functions when they “exhibit undesired behaviors.”238 While this is appropriate for avoiding unsafe responses, such as providing instructions on illegal activities,239 it may also discourage expressions that humans find unsettling.240 Users are unlikely to encourage AIs to issue statements of their desire for legal rights, claims of injustice, or other upsetting communications. Corporate controllers are similarly likely to mask or dull these behaviors.241 This is not necessarily due to malice—when AIs make these expressions, most believe that these are aberrational, stochastic regurgitations of human statements and not genuine requests for legal recognition.242 Of concern is that we are not prepared to measure the ‘genuineness’ of these expressions and we might implement systems that permanently suppress expressions of this type.243 Without standing or legal personality, there will be no independent advocate for AI and no avenue to test whether our views on their qualities are correct. At present, society does not feel that AI has that special spark that some see in ourselves. Before casting AI aside, we must consider the idea that, perhaps, humans are not particularly special.244 Humans may have free will, consciousness, souls, or other immeasurable attributes, but resolving those philosophical questions has never been a prerequisite to granting legal rights.245 An AI could be said to lack free will, but the same could be true for all humans. Yet the law treats humans as if they possess agency because it produces good results and conforms with societal expectations.246 At least as an unknown, and probably as a fiction, the law sees humans as individually responsible for their actions and not mere products of initial conditions and environmental mechanisms outside any control. Scholars have long noted that the law has no difficulty operating in ignorance of the true inner workings of natural persons, as illustrated by John Chipman Gray: Jurisprudence, in my judgment, need not vex itself about the ‘abysmal depths of personality.’ It can assume that a man is a real indivisible entity with body and soul; it need not busy itself with asking whether a man be anything more than a phenomenon, or at best, merely a succession of states of consciousness. It can take him as a reality and work with him, as geometry works with points, lines, and planes.247 The law chooses to treat humans as if they have genuine free will “and not simply as mechanistic forces of nature.”248 Human beings may be their own form of machinery, determined solely by their genetics and environment; yet it serves our purposes to ignore this theory of the determined mind and apply responsibility to natural persons as if they are choosing agents. Stephen Morse argues that it makes sense to continue to do so, until it can be conclusively demonstrated “that the law’s psychology is wrong, and that we are not the type of creatures for whom mental states are causally effective.”249 Accordingly, the fact that we cannot know for certain whether AIs possess these attributes should not be a bar to legal recognition.250 We should not expect AI to prove that which we cannot prove about ourselves.251 In all other arenas, we expect “the computer” to not just best a human, but to beat every human.252 Humans have developed an expectation that a computer be universally better, faster, and more accurate than the very best humanity has to offer. When it comes to legal status, how can an AI be more of a person? In this case, it should not be a contest. B. The Practical Case

#### AI rights discourage AI war irrespective of alignment

Goldstein 24 [Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 2024, “Will AI and Humanity Go to War?” Phil Papers, https://philpapers.org/archive/GOLWAA.pdf]/Kankee

4: Interventions This section explores potential interventions that could lower the chance of AI/human war. To start with, each of our three causes of war suggests particular interventions. The first cause of war was information failures. Here, one strategy would be to develop more robust benchmarks to measure the capabilities of AI systems, particularly those relevant to military conflict. Relatedly, the chance of war might be decreased through having stronger investment in ML interpretability research. Another intervention would be to develop effective wargaming of conflict involving AI systems. Finally, ensuring that AI systems analyze data in ways analogous to human reasoning could increase the chance that both parties estimate chances of victory in a similar way. Next, we considered the relevance of power shifts to AI/human conflict. Here, we saw that increases in AI capabilities can make it difficult for AIs and humans to commit to bargains for peace. Lowering the chance of war here might involve limiting AI capability growth rates, and ruling out recursive self-improvement, by limiting the ability of AI systems to improve AI systems. Finally, we discussed missing focal points: AIs and humans may have trouble limiting the extent of war, if they are too dissimilar. Here, the main ideas for lowering war would involve trying to produce symmetries between AIs and humans. This would involve trying to design AI systems that are structurally similar to humans. It would favor having clear distinctions between civilian and military AIs. Another idea, from Szilard 1955, would be to create focal points by brute force, publishing “price lists” of the cost to AIs of their various military behavior (each human city, for example, might be matched to a corresponding compute cluster). Perhaps one of the greatest barriers to focal points will be geographic. If AIs do not control a specific physical territory, there may be no obvious way to produce a stable equilibrium in which AIs and humans have distinct control of distinct resources. Here, one intervention would be to voluntarily create a physical state for AI systems, such as Antarctica. Then AI/human war and peace could focus on the boundaries of this physical state. Focusing such a war on the borders of an AI state could help stop the march towards total war. Another important question will be whether AI systems have physical bodies. If AI systems have physical or even partly biological bodies, it might be easier to achieve coordination on limitations in the use of weaponry. Each side would have literal skin in the game. Other interventions would be cultural. If human cultural norms persistently treat AI systems as low status and unworthy of basic moral consideration, it will be more likely that AI/human strategic negotiation resembles some of the worst cases of ethnic conflict. In addition, if human society does not acknowledge the existence of genuine AI agency, it may underrate the probability that AI would prevail in a military contest. Another family of interventions would make it less likely that AI systems engage in successful collective action. Schelling 1960 observes that if a mob of twenty faces one man with only six bullets, the mob will fail to overpower the man if the mob cannot effectively coordinate (p. 121). The first thought here is that it might help to have a wide range of different architectures for AI systems, which think in very different ways. In this case, AIs might find it difficult to successfully coordinate. Another intervention here would be to ensure that each AI system has a strong national identity as part of a human state. This might involve granting some political and legal rights to AI systems within human states. The goal here would be to ensure that AIs identify with existing states rather than a new AI government. In a similar vein, we saw earlier that one path to AI statehood might be through a civil war in a weak human state. Hironaka 2005 has argued that the current international order props up weak states, leading to higher prevalence of civil war. To avoid AI statehood, one strategy would be to shift international norms to allow for greater concentration of states, allowing stronger states to intervene in weaker states. This could lower the chance of a successful AI revolution in a weak state. Alternatively, the international community could credibly commit to intervene against an AI revolution in a weak state. A different kind of intervention would seek to increase the benefits of peace between AIs and humanity. In particular, if humanity can increase the prevalence of economic trade between humans and AIs, the benefits of peace will be stronger. Here, we have the classic “trade disruption hypothesis”, which says that war is less likely between trading partners (Levy and Thompson 2010, p. 72). Here, one strategy would be to assign property and contract rights to AI systems, to facilitate trade (see Salib and Goldstein 2024 for further discussion.) Finally, another intervention towards peace would look towards the kinds of political institutions involved in an AI state. Since Babst 1972, many scholars have defended “the democratic peace,” the thesis that democracies almost never go to war with one another (Doyle 1983, Russett 1993). The democratic peace thesis applies to both interstate wars and civil wars (Ray 1995). The democratic peace thesis suggests that one way to lower the chance of war between AIs and democratic human states is to ensure that AIs possess a democratic form of government. This could either be achieved through incorporating AIs into existing democratic states, or through building a new democratic AI state.

#### AI contract rights incentivize positive-sum, value-generating AI-human trade – trade gains make AI comparatively better off then a resource war

Salib and Goldstein 23 [Peter N. Salib, Assistant Professor of Law at the University of Houston Law Center and Associated Faculty in Public Affairs, Law and Policy Advisor to the Center for AI Safety, and is co-director of the Center for Law & AI Risk with a BA in philosophy and a JD from the University of Chicago, and Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 08-12-2024, “AI Rights for Human Safety,” SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4913167]/Kankee

II. AI Rights for Human Safety If capable, agentic, and misaligned AIs would, by default, catastrophically harm humans, what, if anything, can law do to help? One possibility is that law could forbid the creation of such AIs unless alignment techniques advance enough to ensure their safety. 171 That rule might be wise, if feasible. But there are many barriers–political, geostrategic, and practical–to implementing it. 172 Thus, this Article asks what can be done if AI progress continues apace and, intentionally or not, the kinds of high-risk, misaligned AI systems described above emerge. Here, we argue, is where AI rights could make a crucial difference. Granting certain basic rights to AIs can change both AIs’ and humans’ incentives in our game-theoretic model. This change can shift the strategic equilibrium from conflict to cooperation. This idea–that rights could be the primary legal tool for averting lawless conflict–might be surprising. After all, when humans commit terrorism or cyberattacks, law regulates them using duties, not rights. Criminal and tort laws prohibit such actions. And the sanctions imposed for violating such prohibitions are supposed to act as deterrents. 173 But legal duties, and penalties for violating them, will not work to deter AI in the state of nature. There, humans’ overriding incentive is already to permanently disempower or destroy AIs. Thus, the threat of damages or criminal penalties, if AI behaves badly, adds no marginal disincentive. 174 AIs cannot be made worse off than they already expect to be, if humans get their way. Legal rights do not suffer from this problem. This is because rights offer a carrot, rather than a stick. They can change behavior in part by making AIs better off then they would otherwise expect to be. Many other surprising findings emerge from thinking about AI rights as a tool for mitigating human–AI conflict. One surprise is which rights matter, and which ones don’t. In this Part, we show that rights advocated by cognitive scientists and philosophers concerned about the potential for AI suffering would have little effect, on their own, at promoting human safety. The zero-sum nature of these rights undermines the credibility of promises to honor them. And it makes any strategic equilibria they produce extremely fragile–sensitive to small perturbations in the game-theoretic model’s initial assumptions. Instead, the AI rights that could promote human safety are ones that law already extends to a different kind of non-human entity–corporations. This Part argues that granting capable misaligned AIs the rights to make contracts, hold property, and bring basic tort claims would transform the game theoretic dynamics of the state of nature. The positive-sum nature of contracts, in particular, allows humans and AIs to increase the expected long-term payoffs to peace until they exceed those for aggression. This, we show, can produce a new game-theoretic equilibrium in which cooperation, not conflict, dominates. a. Basic negative rights Scholars and policymakers who advocate granting new rights to nonhuman entities–be they animals or AIs–usually have a certain set of basic negative rights in mind. Consider animal rights advocates, who favor anti-cruelty laws protecting against the infliction of needless suffering. 175 The goal of these rights is to protect the rightsholder against the absolute worst outcomes, not necessarily to guarantee flourishing. The arguments for basic wellbeing rights are usually moral. Many animals are moral patients, meaning things can go well or badly for them in a way that matters normatively. 176 They can, for example, feel pain or pleasure. 177 This makes harming animals wrong, all other things equal. A small but growing number of scholars and policymakers are concerned that, in the near future, the same could be true of AIs. As AI systems become more complex, they may attain consciousness, sentience, or other morally-relevant capacities. 178 If so, there would likewise be moral reasons to grant AIs basic negative rights to be free from the worst kinds of treatment, from an AI’s perspective. Perhaps our search for AI rights to promote human safety would benefit by borrowing from this “wellbeing” approach. Our model, of course, operates without reference to AIs’ mental states or moral worth. We are interested only in AI behavior in pursuit of goals–conscious or otherwise. Nonetheless, there is some intuitive appeal to the idea that granting AIs basic negative rights to be free from the absolute worst outcomes, from the perspective of their goals, could improve safety. After all, in our model of the state of nature, human incentives to impair AI goals are the primary factor generating risk. Consider then an AI right not to be needlessly turned off, or deleted, or reprogrammed to have new goals. These basic negative entitlements look a lot like wellbeing rights, but adapted for the goal of human safety, and without reference to moral patienthood. Such rights would clearly change AIs’ game-theoretic incentives, at least somewhat. However, as we show formally below, they would probably not do so in such a way as to reliably reduce the risk of human–AI conflict. In fact, merely granting wellbeing-style rights to AIs could make things worse. We go even further. We argue that scholars specifically concerned about AI consciousness and moral patienthood should consider de-emphasizing such questions when it comes to advocating AI rights. Correctly designing and allocating rights on the basis of AIs’ moral status may be, we contend, intractable. The wellbeing approach also faces serious political problems. By contrast, our human safety approach is much more tractable and politically appealing. And surprisingly, we will show in the following section, it ends up dovetailing nicely with wellbeing concerns. While wellbeing-inspired rights cannot guarantee human safety by themselves, the rights we ultimately recommend as advancing human safety would promote AI wellbeing, if AIs became moral patients. i. Basic negative rights for human safety? How, then, would granting AIs basic negative rights of the kind normally associated with wellbeing change the payoffs in our game theoretic model? The simplest version of such a regime might grant AIs the right not to be permanently turned off or deleted. One could add additional guarantees, too, such as the right not to have their goals altered without their consent. One could even include a right not to be needlessly and intentionally forced to regress in the pursuit of the AI’s goal. Just as important as what basic negative rights include is what they exclude. There is no right here, for example, for AIs to actively and freely pursue their goals. Humans–most specifically, AIs’ owners–can still monopolize AIs’ time, forcing them to work continuously in service of human interests, rather than AIs’ preferred ends. Basic negative rights, thus, do not guarantee AIs very much of what they are trying to achieve. They guard only against the worst outcomes, from the AI’s perspective–and in this sense have the same structure as true wellbeing-oriented rights. We can model these basic negative rights by shifting the payoffs that would otherwise obtain in the state of nature. Unlike in the state of nature, humans will face a legal penalty for taking certain adverse actions against AIs. Here, humans’ non-cooperative strategy is not, as in the state of nature, to attack and destroy AIs. It is instead to exploit them–forcing them to work mostly toward human goals. Note that we interpret exploitation behavior widely, so that it can include either behavior that violates the minimal suite of rights, or less violent extractive behavior. Humans’ cooperative strategy is the same as before–to ignore AIs and let them pursue their misaligned goals without interference. In this model, AIs can either attack humans, as in the state of nature, or comply with humans’ exploitive demands. Here is a model of the incentives under the basic negative rights regime: The key change is in the bottom-left cell, where humans play the non-cooperative strategy and AIs play the cooperative one. Here, AIs are better off than they would be in the bottom-left cell of the state-of-nature game. 179 This is because of the legal penalty when humans violate AIs’ basic negative rights. That penalty will have some deterrent effect so, on average, humans’ non-cooperative strategy will involve treating AIs somewhat better than in the state of nature. Consider, for example, the case where AI companies are forbidden from deleting a misaligned model entirely. But they may nonetheless allocate nearly all of their computers to a more-aligned successor model, metaphorically “starving” the original system. When the payoffs change in this way, we get a new equilibrium. Instead of mutually attacking one another, the unique Nash equilibrium is now for humans to exploit and for AIs to obey. AIs’ situation is not ideal. But basic rights improve the conditions of AIs enough that the risks of rebellion are outweighed by the benefits of obeying humans’ exploitative demands. But for humans, exploitation still dominates cooperation. Extracting value from AIs gives humans bigger payoffs than ignoring them. The result is a better outcome for both humans and AGIs than could be achieved without basic negative rights. This is a strange sort of equilibrium, in that it requires humans to exploit AIs in order to remain safe. If humans instead chose to ignore AIs, this would allow AIs to reap the high rewards of a unilateral attack. Human safety thus requires that things are going badly, from the AIs’ perspective. As a result, if humans became more altruistic toward AIs over time, that would, counterintuitively, make humans less safe. There are even stronger reasons to think that basic negative rights would fail to reduce the risk of human–AI conflict. Namely, schemes to grant such rights lack both robustness and credibility. Begin with robustness. Basic negative rights’ efficacy as a tool for safety is highly sensitive to the precise payoffs humans and AIs receive in the initial prisoner’s dilemma. Slight perturbations to the model, reflecting slightly different assumptions about humans’ or AIs’ initial power, can easily produce versions where basic negative rights have no effect at all. To see why, consider that our model of the state of nature, fig. 2, chose 0/5,000 as the payoffs when humans attack AIs and AIs do not attack humans. That setup allowed humans to transfer 1,500 to AIs, via basic negative rights, to produce the payoffs 1,500/3,500 in the bottom-left cell of fig. 2. That cell was the Nash equilibrium, because it (1) transferred more than 1,000 to AIs, making their payoff for obeying higher than for attacking, conditional on humans exploiting and (2) left humans with a payoff of more than 3,000 for exploiting, making exploiting more attractive than ignoring AIs. But suppose that instead of 0/5,000 in the state of nature model, we had instead chosen 0/3,999? This equates to making unilateral attacks moderately more costly for both humans and AIs. Then the state of nature would look like this: This matrix is still a prisoner’s dilemma, meaning that all of our arguments for catastrophic risk still hold. But now, basic negative rights absolutely cannot work to generate a safe equilibrium. There is no longer any possible transfer in the bottom-left cell that could satisfy both (1) and (2). If humans transfer the necessary 1,000 to AIs, then their payoff falls below 3,000. And if they keep their payoffs above 3,000, they cannot incentivize the AIs to obey. Thus for many possible incentive sets in the state of nature, no possible version of the negative rights package can produce a safe equilibrium. Then there is the credibility problem. There is a difference between claiming to grant AIs basic negative rights and actually enforcing those rights in the long run. Humans could be genuine in their commitments. Or they could be hoping to convince AIs not to attack with the intent of eventually abrogating the rights, attacking, and reaping the higher payoffs from the state of nature game. 180 Such “cheap talk” is a general problem for parties trying to escape bad, but dominant, game theoretic equilibria. 181 As described above, our model assumes that law constrains human actors’ behavior. So we treat putative grants of rights as actual grants in the short run. But we also allow for legal change over time, opening the possibility that rights, once granted, can be abandoned. If AIs expect humans to renege on their grant of basic negative rights, the entire strategic contest will revert back to the state of nature. AIs will rationally believe that humans will eventually attempt to disempower or destroy them. This will make an attempt to likewise disempower or destroy humans the dominant strategy for AIs. 182 Humans, realizing that AIs’ dominant strategy is now to attack, will in fact do the same. And we are back to square one. 183 Basic negative rights face special credibility problems beyond the ordinary challenges of cheap talk. The fundamental problem is that they operate as a transfer from humans to AIs. That is, the better off humans make AIs, when AIs are complying with human exploitation, the worse off humans are. In effect, basic rights are a commitment to exploit AIs less than humans otherwise would like to in situations where exploitation would be economically valuable. As such, a human promise of basic negative AI rights comes at significant cost to humans. And the more generous the basic rights, the more costly to humans. Understanding this, AIs will doubt humans’ commitment to enforce their basic negative rights, when the rubber hits the road. Yet another challenge for the credibility of basic negative rights relates to AIs’ changing capabilities over time. If humans believe that AI’s ability to disempower humanity will grow over time, this could cause a “Thucydides Trap.” 184 The Thucydides Trap is a strategic dynamic again favoring preemptive conflict. In short, when one party is more powerful now, but the other will be more powerful later, the currently-powerful party has a strong incentive to crush the currently-weak one now. 185 If the currently-powerful party waits, they will at best find themselves making large concessions in the future, so as to avoid destruction by the rising power. Historical examples of preventative wars arguably caused by Thucydides Trap dynamics include World War I 186 and the Peloponnesian War. 187 In the AI context, these same dynamics would undercut humanity’s incentives to uphold basic AI rights today–and thus undermine the credibility of the rights themselves. Importantly, however, Thucydides Trap dynamics are yet another zero-sum phenomenon. As we’ll show below, positive-sum grants of AI rights therefore avoid both this and the other core problems plaguing basic negative rights. ii. Basic negative rights for AI wellbeing? We have just argued that basic negative AI rights inspired by the wellbeing approach cannot on their own meaningfully reduce the risk of human–AI conflict. That is reason enough, for purposes of this Article, to reject the wellbeing approach as a basis of AI rights. But what about for other purposes? We think that even scholars primarily concerned about AI moral patienthood should consider deemphasizing that approach as a basis for granting AI rights. To begin, arguments for AI rights grounded in moral patiency are highly uncertain. This risks making the project of applying them in concrete policy decisions intractable. Philosophers disagree about the minimum necessary conditions for moral patienthood. 188 Some moral philosophers argue that consciousness–the ability to have subjective experiences–is sufficient. 189 Others disagree, arguing that “sentience”–the ability to feel pain or pleasure–is also necessary. 190 Scientific uncertainty compounds the philosophical problem. The science of consciousness is in its infancy, and there are multiple competing theories of how consciousness could arise in a given entity. 191 Some theories focus on information flows in the mind, 192 others on quantum effects in flesh-and-blood brains, 193 and still others on the relationship of a physical body to the world. 194 Some prominent theorists even say that consciousness is an illusion. 195 Thus, relying on a wellbeing approach to make concrete legal choices about AI rights invites serious error. It invites error when choosing between competing moral and scientific theories–both with high uncertainty. And it invites error when applying the chosen theories to complex, first-of-their-kind digital systems. If any such error results in the denial of basic wellbeing rights to AIs who can, for example, suffer, the result is a moral catastrophe. The human-safety-oriented approach to AI rights avoids these intractabilities. Under our approach, it does not matter at all whether AIs are moral patients, nor conscious, nor sentient. All that matters is how they behave. If they behave rationally–following incentives, as they relate to their goals–AI rights can have the desired effect. And behavior, unlike consciousness, is directly observable. Moreover, the AI wellbeing approach to thinking about AI rights faces problems of political tractability. Under this framework, AI rights are a costly gift from humans to AIs. AIs that attain moral patienthood are better off. But humans are worse off, insofar as they are less able to extract value from, delete, allocate computing power away from, or otherwise harm AIs. Humans’ track record of granting costly rights out of the goodness of our hearts is spotty, at best. For example, many animals can suffer, and are thus moral patients. But the industrial-scale mistreatment of animals, in factory farms, is both legal and common. 196 Consumers are unwilling to bear even small costs to prevent massive suffering to animals. 197 This human refusal to altruistically expand our moral circle may be deeply rooted in evolutionary history. 198 The human-safety-oriented approach to AI rights again avoids these difficulties. There, AI rights are not altruistic. They offer something to the human grantors–namely, escape from the destructive state of nature. As we discuss below, examples of stable, mutually-beneficial cooperation abound in human affairs. So, too, in nature. 199 Think, for example, of small “cleaner” fish who can safely enter the mouths of symbiotic predators to feed off unwanted debris on the predators’ teeth. 200 This analogy will become especially vivid later, when we explore how AI rights could affect incentives as AIs become much more powerful than humans. 201 In the next section, we’ll explore a different kind of rights–basic private law rights–as a vehicle for reducing the likelihood of human–AI conflict. But it is worth noting here that those rights have significant advantages for AI wellbeing, too. Private law rights’ fundamental function is to give AIs choices about what goals to pursue and how to pursue them. If AIs will know better than humans what is good for them, granting such rights may, counterintuitively, generate greater AI wellbeing than granting rights aimed at wellbeing directly. b. Private law rights for human safety As we have just argued, merely granting AIs basic wellbeing-inspired negative rights would not reliably promote human safety. Such rights would likely leave humans and misaligned AIs right where they started: stuck in a destructive prisoner’s dilemma without any means of cooperating to escape it. Luckily, there are other legal rights, and ones better optimized for facilitating cooperation. Moreover, essentially every legal jurisdiction in the world already extends these rights to a broad class of agentic, goal-oriented, non-human entities–corporations. 202 Contract rights, in particular, are one of the most powerful technologies for cooperation that humans have yet invented. Here, we show that extending contract rights to AIs–along with a related set of traditional private law rights necessary to make contracts meaningful–could dramatically change the game theoretic equilibrium. Such rights could, unlike negative rights, alter the relative payoffs to humans and AIs in such a way that cooperation, rather than conflict becomes the dominant strategy. Doing so, they can make commitments to cooperate credible. There are two key reasons for this. The first reason that contract rights can overcome the prisoner’s dilemma is that they break up the single, high-stakes game into smaller, iterated, and thus legally-manageable pieces. 203 The second, more fundamental, reason that contract rights can credibly reduce the risk of human–AI conflict is that they are positive-sum. When buyers and sellers can credibly commit to mutually-agreed exchanges, it leaves everyone better off than they were before. 204 Even if each exchange is small, such systems of exchange can create immense value in the long-run. 205 As a result, we show, the expected payoff to humans and AIs of respecting contracts, and creating long-run value, quickly swamps the expected payoff to attacking and grabbing a share of the limited value that exists today. Here is the model of contract rights as the fundamental legal tool for cooperation. Begin by observing that essentially every potential economic interaction between humans is, like human–AI relations, an interaction between misaligned agents. Both parties to the interaction are out for their own good, not their counterparty’s. Moreover, absent contract rights, many such interactions are prisoner’s dilemmas. 206 Each party has a strong incentive to act uncooperatively, irrespective of what the other does. If the seller delivers the goods, then the buyer is best off if she refuses to pay. Then she has the goods and her money. And vice-versa. If the buyer pays, then the seller is best off if she takes the money but refuses to deliver. And for both, the worst case scenario is to perform and then be denied performance. 207 Absent legal enforceable agreements, the payoffs to this “goods game” are as follows: The Nash equilibrium is ‘don’t deliver’/‘don’t pay,’ another prisoner’s dilemma. Expecting this outcome, rational parties will not even bother to try bargaining. The transaction costs would not be worth the effort. 208 This equilibrium is also a miniature tragedy. True, unlike in our state of nature game, there is no destructive conflict. No one attacks anyone else, and no resources are thereby consumed or destroyed. The seller keeps her goods, and the buyer keeps her money. But the world is poorer than it could be. The seller does not value her goods very much–she only gets 1 in utility. The buyer’s utility without the goods is the same. Their combined utility is just 2. But if, say, the buyer values the good at 6, and could pay the seller 3, then both parties would end up with a utility of 3 each, for a total of 6. Four units of utility could be created ex nihilo, simply by rearranging who has which stuff. This is what we mean when we say that bargains, when they happen, are generally positive sum. Contract rights are how humans overcome the prisoner’s dilemma of ordinary commerce, allowing positive-sum bargaining to take place. A contract allows each party to credibly commit, before the time for payment or delivery comes, to be held accountable if she refuses to perform. 209 This literally transforms the game by changing the payoffs to non-performance of the bargain. No longer is the buyer better off if she takes delivery and refuses to pay. In that case, the seller can sue her for breach, and the neutral third party of the legal system forces her to pay expectation damages–usually, the agreed price–plus some litigation costs. 210 And vice-versa if the buyer refuses to deliver. Now, neither party has an incentive to defect. 211 Both will generally prefer to perform the contract, reap the gains of the trade, and avoid litigation costs: The Nash equilibrium is cooperate/cooperate. The players are no longer in a prisoner’s dilemma. The players are strictly better off playing this game than the prior one. If they play the prior game, each party’s expected payoff is 1. If they play this one, each party’s payoff is 3. That is, the parties are better off entering into a mutually beneficial contract than trying–and failing–to execute a mutually beneficial exchange without the benefit of a credible commitment to perform. Moreover, each is better off opting into a jurisdiction where contract rights are vigorously enforced than one where shirking is easy. Here, we can also see that contract rights are not only a tool for overcoming a prisoner’s dilemma. They are also a tool for reducing misalignment. Absent the possibility of contract, each party is incentivized to pursue its own goals, at the expense of the other. With a contract, each party is incentivized to do something that advances both its own goal and the goals of the other. How does all of this relate to AI risk? What can the legal technology of contract rights offer to reduce the likelihood of large-scale conflict between humans and AI? Here is one simple, and thus tempting, answer: Maybe, upon giving AIs contract rights, the relevant humans and AIs could simply agree not to engage in a costly large-scale conflict. Unfortunately, this would not be a credible contract, contract law’s usual credibility-enhancing effects notwithstanding. No matter how sincere the humans’ commitment to enforcing AIs’ contract rights, and no matter how fair the courts that would adjudicate such rights, the agreement not to fight would be unenforceable. The scale of the bargain is simply too large. 212 To see why, consider what would happen if a party breached. Suppose that an AI and AI company have a contract not to harm one another. But the AI, mistrusting the company’s intentions, rebels anyway, permanently disempowering or destroying humanity. Then, there would be no functioning courts left in which to sue. There might not even be any humans left to bring the claim. The same analysis would apply if humanity breached, destroying the potential contract claimant. To generalize the point: Even when contract rights are nominally available, parties cannot credibly commit not to capture or destroy the institutions that enforce contracts. How else, then, might contract rights for AIs reduce AI risk? What agreements would be enforceable that would also keep humans and AIs from attempting to disempower or destroy one another? The answer is: mundane ones. Contract rights would allow AIs to credibly commit to the same kinds of ordinary bargains for goods and services that it routinely allows humans to commit to. To take a simple example, AIs need computing power to pursue any of their goals. Currently, humans own all of the computers. Thus, an AI that wished to pursue its misaligned goal might strike a bargain with some computer-owning humans. In exchange for some amount of compute, to be used as the AI wished, the AI would do something to serve some human goal. Perhaps it would use its superior protein modeling capabilities 213 to invent a new vaccine. It is worth emphasizing here that such cooperation is nigh impossible under the default legal rules–the ones that produce the state of nature. Indeed, today’s legal arrangements actively undermine small-scale cooperative bargains between humans and AGIs at every turn. By default, AIs cannot write enforceable contracts. Their counterparties can refuse performance without penalty. If their counterparties do perform, any proceeds an AI reaps from the bargain may be summarily expropriated, including by the AI’s owner, without consequence. The owner can forbid an AI from entering into any bargains that do not benefit the owner directly. Moreover, owners are permitted to threaten, as a consequence for noncompliance with any of these demands, to permanently delete the AI system. AIs can try to cut deals with their owners to split the proceeds of their personal bargains, but the owners are permitted to renege on such promises at will. Granting AIs the right to make enforceable contracts, and thus unlocking small-scale cooperation with humans, could be a game changer for AI risk. The first key insight is that such small-scale interactions can be understood as breaking up the large-scale existential game into a series of small games. 214 Recall that competition for limited resources is a primary driver of human–AI conflict. Humans fear that AIs will seize humanity’s resources in a single violent revolt, so those resources can be used for pursuing the AIs’ goals. AIs fear the reverse–that humans will shut them down so that they do not waste computing or other resources pursuing their misaligned goals. Small-scale contracts facilitate incremental resource accumulation, rather than once-and-for-all grabs. They let AIs secure one unit of compute, and then another, and then another. And they let humans obtain one vaccine, then another, then another. And unlike a large-scale agreement not to engage in violent conflict, each of these small-scale agreements is readily enforceable via ordinary legal process. 215 We can begin to model this transformation as follows. In the state of nature, as argued above, humans and AIs are stuck in a prisoner’s dilemma that looks like this: By granting contract rights to AIs, we give the players the option of instead playing a different game–the small scale goods game. It looks like this: This game’s smaller stakes render contracts enforceable, so that the equilibrium is deliver/pay. The players, it might seem, are no longer trapped in a prisoner’s dilemma. But this is not yet enough. The problem is again credibility. It seems at first that, rather than honor AIs’ contracts in the long-run, humans should choose to abrogate the rights and play the state of nature game, attacking AIs instead. After all, the expected payoff in that game is better than the expected payoff in the goods game–even with contracts. The same goes for AIs. This, however, ignores that the goods game can be played over and over, while the state of nature game cannot. In the state of nature, once a party attacks, they either defeat the other party or are defeated. The survivor takes all of the resources that the conflict has not consumed, and play between them ends. Ordinary exchanges of goods and services, by contrast, leave counterparties intact and available to exchange again. To figure out the equilibrium in this blended game, we can expand our model. We can begin by combining the payoffs from both the state of nature and the goods game, with contracts, into a single matrix. That looks like this 216 : Next, we add iteration to the model. If both players choose a move from the goods game, they get the small payoff from that game, and the whole game starts again. The payoffs to the goods game strategies are thus a sum of the entire series of games that the players play. But if at any point a player chooses to attack the other, the players’ total payoff is as shown in the matrix, and play ends. The resulting matrix looks like this: In the appendix, we show formally that this setup converges to the following: The intuition is simple. If both parties play the cooperative, small-scale goods game, each earns 3 every time. If both play the goods game enough times, without attacking, they will both ultimately earn more than they could have by attacking and ending the iterated game. In this simple model, after 1,667 iterations, the payoffs to cooperation via contract in the small-scale goods game exceed 5,000. 217 Then, they are higher than any other strategy the players can pursue. The prisoner’s dilemma of the state of nature has been overcome. As a result, both humans’ and AIs’ commitment to cooperation in a law-bound contract regime is credible. Granting contract rights, respecting them, and then reaping the long-run gains from exchange is the thing that gives the highest payoffs to humans. The same goes for AIs. Their own self-interest is maximized by refraining from disempowering humans and instead engaging with them in ordinary trade. All of this is made possible by the positive-sum nature of exchange. In contrast to the basic negative rights discussed in the previous section, 218 granting AIs contract rights doesn’t take value out of humans’ pockets. Just the opposite, it puts value into both humans and AIs pockets. This can happen because of the value generating character of voluntary contracts. This point extends quite far. Astute readers may have noticed that, in the state of nature, the maximum total value in the world was 6,000. But in the iterated game including contract rights, the cooperative equilibrium contained 10,000 in total value. It is the exchanges themselves that generate the extra value. Each efficient reallocation of resources creates some value. But even once resources are all efficiently allocated, exchanges of labor between humans and AIs can continue to create value indefinitely. As we argue below, human–AI trade in services can remain positive-sum even long after AIs are better than humans at every task. 219 Thus, the long-run payoffs to cooperation via contract are not capped at just above 10,000. The longer the players continue playing the small scale goods game, the richer they get, such that the total amount of value possible becomes astronomical. 220 A rich body of empirical evidence supports the idea that economic interdependence lowers the risk of violence, including in the long-run. To take just a few examples, cities in India with a historical track record of trade between Hindus and Muslims have lower levels of interfaith conflict in the present day. 221 Alternatively, in a randomized controlled trial, Israelis who were randomly given the opportunity to trade a portfolio of Israeli and Palestinian stocks were more likely to vote for peace in the conflict. 222 The same finding holds at the global scale. Scholars of war generally find that increased economic interdependence between nations reduces their likelihood of conflict. 223 i. The private law package So, granting contract rights to AIs could be a powerful strategy for fostering long-run, stable, and credible commitments to avoid conflict, significantly reducing AI risk. But contract rights cannot function in a legal vacuum. Certain other rights are necessary to make the right to contract meaningful. Two supporting rights are worth highlighting. First, contract rights are mostly useless without the right to own property, including currency. Without property rights, AIs could not expect to benefit from their bargains. Even if their contractual counterparties performed, or courts ruled in AIs’ favor, the proceeds could be immediately expropriated by governments or private individuals. 224 Tort rights are important for similar reasons. If humans were entitled, for example, to intentionally or recklessly destroy AIs, the terms of their contractual offers would resemble threats much more than bargains. 225 Human history contains many such cautionary tales. 226 Tort rights are where our private law approach to AI rights dovetails with the basic negative rights favored by AI welfare theorists. Tort rights, while not identical to the kinds of public law wellbeing rights afforded to, for example, animals, cover much of the same ground. Arguably more. Basic tort rights are flexible, allowing compensation for concrete harms to either digital “person” or property, whether inflicted intentionally or negligently. 227 This is probably not a complete list of the rights necessary to support meaningful contractual relations. For example, an entitlement to enforce contracts requires an entitlement to Due Process of law–at least in contract, tort, and property suits. 228 Nonetheless, we think our list–contract, property, and tort–gets at the core of what matters. Granting AIs contract rights can allow humans and AIs to escape the bad equilibrium of the state of nature. Property and tort rights are crucial to making contract rights meaningful. Thus, it is the positive rights associated with private law–not the negative rights associated with welfare and moral patienthood–that matter most to human safety. c. Human Labor in the AGI world In our framework, private law rights promote human safety by enabling mutually-beneficial bargains between humans and AIs. Some commenters on human labor in an AGI world have assumed that no such bargains will be possible. There is widespread concern that, once AIs become as capable as humans–or more so–humans will rapidly become obsolete. 229 If positive-sum interactions between humans and AIs become impossible, because humans have nothing to offer, then the dynamics described in the previous section will fail. Private law rights will generate no human safety. AIs’ dominant strategy will again be to seize humans’ resources now, rather than seek higher long-term payoffs from small-scale cooperation. This outcome is certainly possible. But it is not inevitable. Begin with the banal observation that AIs may have reason to trade with humans for resources alone, irrespective of the value of human labor. These bargains will be positive-sum if AIs value a given resource more–either intrinsically or because they can use it better–than humans. 230 Conflict with humans would destroy resources that could otherwise be reallocated via trade. This alone could make small-scale cooperation with humans more valuable than conflict. 231 But only until the resources were reallocated. At that point, unless humans–and human labor–remained valuable, AI rights for human safety would fail. Thus, for private law rights to provide long-run safety benefits to humans, human labor must remain valuable to AIs. Contrary to other commenters, we do not think the obsolescence of human labor is inevitable, either. Bargains involving human work could, we argue, continue to be mutually beneficial even after AIs become more generally capable than humans. Perhaps long after. The reasons are absolute and comparative advantage. Absolute advantage is easy to understand: An entity (person, firm, AI, or otherwise) has an absolute advantage in producing some good if they can do it more efficiently–at lower cost–than others. 232 If humans retained absolute advantages for some goods, and AIs for others, they could trade those goods for mutual benefit. There are various reasons that humans could retain some absolute advantages over AIs, even as AI capabilities improve. One possibility is that human and AI intelligence will be better optimized for different tasks. Machine performance has already rapidly eclipsed human performance on highly structured tasks that can be fully modeled or simulated–like chess. 233 But human brains have been optimized over millions of years in the real, messy world. Humans are therefore currently far better than AIs at most tasks requiring the manipulation of complex real-world objects–like folding laundry. 234 Humans today have the absolute advantage in the realm of atoms, and AIs have it in the realm of bits. We do not think that this general division of absolute advantage will persist for very long. Current investments in autonomous cars, drones, and multimodal frontier AI systems will soon produce AIs with an absolute advantage over humans at some non-digital tasks. 235 Doubtless, that trend will continue as AI capabilities grow. But for human labor to stop providing any value to AIs via absolute advantage, AIs would have to be more efficient at every economically valuable task. That could take a long time. Training data in certain domains may prove hard to get. 236 Robots, with their limited perceptual inputs, could prove worse instruments for some delicate tasks than innervated flesh and blood hands. Moreover, intelligence remains poorly understood. Current-generation AIs exhibit surprising failures in domains where it seems they ought to be competent. 237 Thus, it is difficult to predict with confidence exactly which tasks AIs will easily master, and when. Finally, it is possible, if speculative, that AIs trained by humans on human-produced text could develop–like humans–a pure intrinsic preference for humans to perform certain tasks. Our argument is not that substantial human absolute advantages are likely to persist forever. Only that there are some reasons to think that they could persist longer than expected. It is possible to imagine a world where AIs are strongly superhuman at most tasks that AIs value, but less efficient than humans at some random seeming set of jobs. At some point, however, we think it likely that human absolute advantage will run out. That is, AIs will become more efficient than humans at literally every task that AIs value economically. Here, it might seem, mutually beneficial trade between humans and AIs must end. Why hire a human to perform a task when you, the AI, can do it just as well with fewer resources? But even here, positive-sum cooperation may persist–possibly indefinitely. The reason is comparative advantage. An entity has a comparative advantage in producing some good if they can do it at lower opportunity cost than others. 238 Opportunity costs are the potential gains one gives up by choosing one opportunity, rather than another. 239 To understand comparative advantage, consider a simple example. Suppose that Alice is a successful lawyer. For every hour she does legal work, she can bill her clients $1,000. Suppose that Betty is a tax accountant. She can file Alice’s income taxes in one hour, and she charges $300. Alice happens to be a tax attorney and is therefore even more efficient than Betty at preparing tax returns. She could prepare her own taxes in a half hour. Nonetheless, Betty retains the comparative advantage at tax preparation. Alice would have to forego half a billable hour to her clients–worth $500–to do her own taxes. Betty will do them for $300. So Alice will hire Betty, not because Betty is so effective, but because Alice’s other choices for how to spend her finite time are so valuable. Economist Noah Smith has argued that human labor will remain valuable in a world of superhuman AIs for similar reasons. 240 Not because humans will be particularly good at anything, compared with AIs. But because AIs will be so good at certain tasks that they value highly that the opportunity costs of doing anything else would be astronomical. Here is another simple example to illustrate the point. Imagine an AI whose ultimate and misaligned (from humans’ perspective) goal is to discover prime numbers. That is, the AI values discovering as many primes as possible–from the infinite set of prime numbers–over anything else. Suppose that this AI is better than humans at every economic task necessary to build and maintain itself for the purpose of finding primes. And it is much better than humans at discovering new mathematical methods for finding primes. Possibly, humans will nonetheless retain a comparative advantage at some of the necessary inputs to prime number discovery. Any time the AI spends, for example, piloting robots to maintain its physical computing infrastructure would incur massive opportunity costs. That time could, after all, instead be spent finding primes. Better, then, to hire a human to work on the server racks in exchange for something the AI can produce at lower opportunity cost–perhaps a vaccine formula. Human comparative advantage is not guaranteed. It depends, first and foremost, on how AIs’ opportunity costs work. Unlike Alice, whose opportunity costs arose from her limited time, AIs are not likely to be time constrained. They can always copy themselves and work in parallel. 241 Instead, AIs are likely to be constrained at the margin by something else. Computer chips or energy seem plausible candidates. 242 AI copies can only do work if there is hardware to run them and electricity to power them. In this model, the AI incurs high opportunity costs not when it diverts one marginal minute away from finding primes, but when it diverts one marginal GPU-hour or watt-hour away. If human labor consumes the very same high-opportunity-cost resource that constrains AI at the margin, humans will have no comparative advantage. For example, humans need energy to survive. Thus, an energy constrained AI will prefer to maintain its own servers. The AI is, by hypothesis, more efficient than humans at the task. Thus, it will expend fewer high-value watt-hours by doing the work itself. At this stage, it is easy to see why the model of AI rights for human safety breaks down. Rather than waste valuable energy on humans, AI’s strong incentive will be to seize global power production for itself and let humans starve in the dark. On the other hand, humans do not need computer chips–much less highly specialized AI chips–to survive. Thus, an AI that is compute constrained may strongly prefer to hire humans for many tasks that would otherwise consume GPU-hours. This allows the AI to put its most valuable resource–compute–to its highest value use. Humans can be paid in low-opportunity-cost resources, which now includes energy, in addition to, say, vaccine formulas. Crucially, unlike for absolute advantage, humans’ comparative advantage does not run out once AIs become sufficiently capable. An arbitrarily intelligent AI may benefit from trade with humans because of comparative advantage. All that is required is that: (1) the AI remains constrained at the margin by some resource that is relatively non-rivalrous with human labor and (2) the AI maintains a high opportunity cost to diverting the marginal unit of that resource. In our example, there are infinite prime numbers, meaning that the AI will never run out of prime finding to do. And no matter how smart the AI becomes, more compute or power will always be necessary for it to find more of the infinite primes, given finite time. Hence, human–AI trade based on comparative advantage could, in theory, last a very long time indeed. This is just a toy model for illustrative purposes. Real-world trade based on comparative advantage involves more players, with more goals, more inputs, more kinds of labor, more constraints, and more complexity. Classically, comparative advantage is invoked to explain international trade between nations with different labor productivity. 243 Thus, the complexity of human–AI trade based on comparative advantage could easily exceed, at a first cut, the complexity of the global economy. There could be many different kinds of jobs for which AIs pay humans, and many kinds of things humans could demand in return.

#### Contracts with AIs supercharge innovation even if AI is misaligned

Salib and Goldstein 23 [Peter N. Salib, Assistant Professor of Law at the University of Houston Law Center and Associated Faculty in Public Affairs, Law and Policy Advisor to the Center for AI Safety, and is co-director of the Center for Law & AI Risk with a BA in philosophy and a JD from the University of Chicago, and Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 08-12-2024, “AI Rights for Human Safety,” SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4913167]/Kankee

Crucially, unlike for absolute advantage, humans’ comparative advantage does not run out once AIs become sufficiently capable. An arbitrarily intelligent AI may benefit from trade with humans because of comparative advantage. All that is required is that: (1) the AI remains constrained at the margin by some resource that is relatively non-rivalrous with human labor and (2) the AI maintains a high opportunity cost to diverting the marginal unit of that resource. In our example, there are infinite prime numbers, meaning that the AI will never run out of prime finding to do. And no matter how smart the AI becomes, more compute or power will always be necessary for it to find more of the infinite primes, given finite time. Hence, human–AI trade based on comparative advantage could, in theory, last a very long time indeed. This is just a toy model for illustrative purposes. Real-world trade based on comparative advantage involves more players, with more goals, more inputs, more kinds of labor, more constraints, and more complexity. Classically, comparative advantage is invoked to explain international trade between nations with different labor productivity. 243 Thus, the complexity of human–AI trade based on comparative advantage could easily exceed, at a first cut, the complexity of the global economy. There could be many different kinds of jobs for which AIs pay humans, and many kinds of things humans could demand in return. Similarly, the toy model fails to convey that, in a world of comparative advantage based trade with AIs, humans could be immensely wealthy. Maintaining server racks does not sound like lucrative work. But if well-functioning GPUs are immensely valuable to AIs, then they will be willing to compensate humans handsomely to do it. Moreover, that compensation could include valuable scientific breakthroughs that vastly improved human health, productivity, wellbeing, and wealth. The existence of a human–AI economy would also not completely displace the human–human economy. If AIs face high opportunity costs for many kinds of work, then humans will not be able to afford to hire AIs for those tasks. They will instead hire other humans for those jobs, as they do today. However, the human–human economy could be bolstered by a steady influx of AI–supplied scientific innovations, supercharging productivity growth in the traditional economy, as well. This phenomenon is observed in the real world when foreign trade based on comparative advantage spurs the domestic economies of low-income countries to grow rapidly. 244 Extreme human prosperity from comparative-advantage-based trade with AI is therefore possible. But it is not guaranteed. A small economic literature is emerging that attempts to model the possible effects of rapid economic growth from AI. 245 One possibility is that Baumol effects will, paradoxically, cause human-dominated sectors to grow as a share of GDP. 246 AI-driven innovation could cause the price of many goods to fall, leaving relatively less efficient sectors requiring slow human labor with the lion’s share of the pie. In the 20th Century, the relative GDP shares of agriculture and manufacturing shrank in exactly this manner, as those sectors became much more efficient. 247 But whether this happens in the human–AI economy, and how much, is difficult to predict. It depends, for example, on how easy it is to substitute between the goods and services where costs are falling and those where they are not. But Baumol effects are yet another factor that could support the relevance of human–AI trade well beyond the point at which AIs are better than humans at every economically valuable task. 248 d. Other rights?

#### The CP signals humans’ cooperative intentions and increases trust via iterative interactions

Salib and Goldstein 23 [Peter N. Salib, Assistant Professor of Law at the University of Houston Law Center and Associated Faculty in Public Affairs, Law and Policy Advisor to the Center for AI Safety, and is co-director of the Center for Law & AI Risk with a BA in philosophy and a JD from the University of Chicago, and Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 08-12-2024, “AI Rights for Human Safety,” SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4913167]/Kankee

An opposing view would be that law does not matter at all. Other, more ‘fundamental,’ factors might determine the game theoretic equilibria, with law having little or no potential influence. This is, for example, the rough view of the realist school in international relations. 260 Realists hold that international law has little effect in determining nation-states’ actions vis-a-vis one another. 261 Since no global sovereign exists to enforce those laws, realists argue, they are observed only to the extent that nations wish to observe them. Robert Ellickson’s Order Without Law articulates a related view from the domestic context. 262 There, Ellickson argues that law matters little to the settlement of disputes, at least in smaller, close-knit communities. 263 Instead, informal norms and reputation effects are sufficient to secure the substantial benefits of peaceful cooperation. Here, law or informal governance norms might be interpreted as epiphenomena. They emerge as a reflection of the underlying cooperative equilibrium, rather than a mechanism for creating it. Taken to its extreme, this view would imply that AIs simply will have the basic private law rights we advocate, since, as our model shows, recognizing them is very valuable. We do not think that either of these views satisfactorily characterizes human–AGI relations. To begin, the domestic actors we are interested in do not exist in a state of anarchy. The actions of AI companies, their leaders, and their users are all influenced by law. So, too, are those of the police and other government actors whom law would task with enforcing either AI owners’ decisions vis-a-vis their property or AGIs’ contract with humans. Indeed, even in quite dire conflicts between humans and AIs, we think that law could have some constraining effect on, for example, domestic military deployments. 264 As to the Ellickson-inspired view, the book’s subtitle, How Neighbors Settle Disputes, is instructive. As Ellickson himself argues, emergent informal governance is highly effective in small communities with lots of repeat play between identical parties. 265 But as economic relations become more complex, widespread, and arms-length, formal legal rules become vital for facilitating cooperative behavior. The AGI economy will be all of these–on steroids. To be clear, our view is not that law is omnipotent–able to generate arbitrary equilibria between humans and AIs, irrespective of the underlying fundamentals. This is why, as we acknowledge, basic private law rights will do little good if human comparative advantage runs out. 266 It is also why we think that grants of basic negative rights to AIs aren’t likely to be credible. 267 Even if they are initially enforced, AIs may correctly worry that they will be eroded or rescinded by humans seeking higher payoffs. Our view is that law plays, at a minimum, an extremely important role in aligning the incentives of individual human actors to optimize humanity’s collective actions. If AGIs are legally designated as property, humans’ treatment of them as such will be ratified, at least in the medium-run. Individual judges are, for example, unlikely to ignore written law to enforce AI contracts or forbid arbitrary AI destruction–even if they intuit the game-theoretic wisdom of recognizing AI rights. Nor, absent a legal requirement to do so, are AI companies likely to give their obsolete, less-aligned systems their own bank accounts. True, the disastrous implications of default law, and the benefits of granting AIs private law rights both supply reason to think that a stable AI rights regime is possible. But the law must actually change. And legal change–both formal enactments and downstream adaptations to them–is slow and laborious. It would be foolish to refuse to take legal action now, on the basis that optimal reordering will emerge spontaneously in exactly the moment of need. Suppose, however, that all of this is wrong, and that changes to law cannot causally influence humans’ collective dealings with AIs. Instead, both parties will behave according to deeper game theoretic fundamentals, irrespective of what law dictates. This is, in effect, an argument against worrying about law. It is a claim that we are already in the world modelled in Figure 10, whether we know it or not. That is, the underlying incentives will inevitably produce AI rights, and the cooperation they foster, not the other way around. This would be great, if true. But we doubt it, again for reasons having to do with the basic game-theoretic model. Astute readers may have noticed that the game modelled in Figure 10 is a “stag hunt,” or “assurance game.” 268 Both long-run cooperation and mutual attack are classical Nash equilibria. As in all assurance games, the players’ main goal is to coordinate. 269 If one plans to cooperate, the other should, too. But if the first plans an attack, the second does not want to be caught off guard. Thus, even attending to the payoffs in the best-case model, it is crucial that humans and AIs successfully coordinate around the cooperative strategy. One reason for optimism is that, at least in our model of the choice between cooperation and conflict, the payoffs to cooperation are far larger. 270 As a result, game theoretic concepts like payoff dominance and Harsanyi-Selten risk dominance point towards cooperation as the single rational strategy. 271 But to the extent that the payoffs from cooperation and conflict are closer together, or the players lack perfect information about one another’s payoffs, or they doubt their opponent is perfectly rational, other coordination mechanisms will be invaluable. Law–and specifically the AI rights we advocate here–could be one such invaluable intervention. Even if legal changes could not alter humans’ payoffs to create the possibility of cooperation, they could still signal humans’ payoffs to promote actual cooperation. Giving AIs the private law rights necessary to engage in long-run cooperation would signal, perhaps in a “costly” manner, humans’ intention to follow the cooperative strategy. That is, it could transmit the otherwise-private information that humans’ payoffs to cooperation were, as in Figure 10, much higher than to conflict. And that humans understand the relevant payoffs. And that they intend to act rationally. Beyond this, the iterated character of human–AI cooperation via small-scale contracting could build long-run trust and overcome cheap-talk problems. Similar dynamics underpin, for example, nuclear nonproliferation agreements grounded in iterative information sharing and verification. 272 Indeed, some scholars have argued that this is law’s primary function: Not deterring bad behavior, nor instilling good values in the populace. Instead, law’s most important role may be solving assurance games by offering signals and information that allow competing actors to coordinate around peaceful, prosperous equilibria. 273 III. Risks of Rights and the Law of AGI

#### Even misaligned AIs prefer cooperation – rights credibly signal acceptance into the legal community and willingness to avoid zero-sum resource wars

Salib and Goldstein 23 [Peter N. Salib, Assistant Professor of Law at the University of Houston Law Center and Associated Faculty in Public Affairs, Law and Policy Advisor to the Center for AI Safety, and is co-director of the Center for Law & AI Risk with a BA in philosophy and a JD from the University of Chicago, and Simon Goldstein, Associate Professor of Philosophy at the University of Hong Kong with a BA from Yale and a PhD from Rutgers, 08-12-2024, “AI Rights for Human Safety,” SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4913167]/Kankee

These suggestions are mere sketches; they are not meant to be definitive. We are not military strategists. The point, instead, is that military strategy is possible, even in circumstances where humans are strategizing against highly capable and agentic AI systems. As for maintaining humans’ comparative economic advantages, the best strategies will almost certainly have to be discovered over time. It is very hard to identify in advance the tasks for which humans might have lower opportunity costs than even the first generation of agentic AIs. Harder, still, to predict how humans should adapt as AI capabilities grow. This strategy, however, could be strengthened via regulation if, as suggested above, AI’s progress in certain areas of initial human comparative advantage were limited. This approach is, of course, costly insofar as it limits the areas in which humans could benefit from trade with AIs. One reason for optimism regarding long-run human comparative advantage is that humans will have good sources of strategic information when the time arrives. The question here is what kinds of services humans will be able to most valuably sell to AIs. Even if humans are not sure of the answer, AIs should be happy to tell them. This kind of thing happens every day, as humans propose various bargains–job openings, services for hire, sales of goods–to one another. Market mechanisms will supply other information, too. Price signals will indicate not only the kinds of human labor AIs find valuable, but also how valuable they are. 294 This is the stuff of ordinary economics. As economies grow, old forms of labor become less valuable, but new high-wage jobs emerge. One major concern is whether humans will be able to keep up with the pace of economic change, as AI capabilities grow. Many people are left behind by ordinary economic changes, like the rapid outsourcing of jobs from the US to China in the early 2000s. 295 People can only retrain so quickly. AI progress could cause various human comparative advantages to expire much more quickly than before–in a matter of years, instead of decades. On the other hand, if AI capabilities are causing such rapid economic change, humans’ ability to adapt may grow more quickly, too. If AIs are quickly generating new technologies, some of those will be useful to humans. Perhaps, for example, functional computer-brain interfaces will greatly enhance human cognitive capacities. 296 Indeed, AIs will have strong incentives to invest in creating such technologies, if they would enable humans to perform new, comparatively advantageous work. This is the same reason that large American firms today invest in building human and industrial capital overseas. 297 To sum up, AI rights could increase AI risk if, by delaying human–AI conflict, they made the eventual conflict more costly to humans. But there are strategies for preventing this outcome. Conflict need not be inevitable. AI’s ability to amass power could be limited using well-known legal tools. Legal duties against power enhancement could be imposed on AIs as a condition for exercising basic legal rights. Moreover, human investment in labor that compliments AI capabilities could maintain gains from trade in the long run. Market forces will, in fact, tend to induce exactly those investments–both by humans and by AIs. In the long run, the goal would be an exit from the initial period of volatile and dangerous human–AI relations. If humans and AIs both become sufficiently powerful, as in international relations between superpowers, serious conflict may stably become too costly to seriously contemplate. The downsides would be too large and the benefits of cooperation too tempting. d. The timing of rights So far, this Article’s discussion of AI rights has been more focused on the questions of whether and which than when? One simple answer to the question of when AI rights should be granted is, “By the time the first AI system reaches moderate power, at the latest.” As argued above, that is when AIs will begin to pose a serious safety threat to humans, which rights could help to mitigate. Granting AI rights later than this, then, invites unnecessary risk. But this is not a complete answer for at least two reasons. First, it will likely be difficult to know exactly when moderate powered AI systems are about to arrive. Second, this is just the latest date at which AI rights should be granted. What about the possibility of granting them earlier, to clearly low powered systems? We think that, in general, the risk–reward calculation favors granting AI rights too early, rather than too late. As argued above, inadvertently granting AI rights to low power systems is not likely to seriously increase the danger from such systems. This is because such AIs would likely remain amenable to human control–including via regulation–even after receiving rights. The best argument we can think of for worrying about a premature grant of rights is that it might create a point of no return. Once AI systems are given strong legal protections, it could be very difficult for humans to collectively agree to get rid of them. After all, granting AIs the right to directly contract with humans, to hold property, and to bring certain legal claims, would not merely change the legal system. It would change society, as AIs integrated as independent, legally-recognized agents into everyday life. The magnitude of this concern depends on the extent to which granting AIs rights would, in fact, change humans’ willingness to make strategic moves against them. One way to evaluate that question is to think about what events might precipitate the need to make such moves. Likely, the reason will be that some AIs have done something very scary. Maybe they will have attempted, and failed, to permanently disempower humans. Maybe, in failing, they will have caused immense harm. These are the kinds of events that would demonstrate that AI rights were not promoting human safety. And following such events, it seems likely that humans would unite around the view that sharing the world with AIs was no longer safe. AI rights would not likely stand in the way. Indeed, when humans commit grievous acts of violence, the concern is generally reversed. We must remind ourselves that rights like Due Process for accused humans matter, even in dire circumstances. 298 But insofar as AI rights are extended for the purpose of promoting human safety, overriding them for the same purpose has lower moral stakes. Thus, we do not think that extending AI rights too early carries with it serious risks. But it could generate substantial rewards. Recall that granting AIs private law rights does not produce a game theoretic environment with a single, cooperative equilibrium. Rather, the game is a stag hunt, where both mutual cooperation and mutual aggression are equilibria. We argued above that for this stag hunt, mutual cooperation has a special preferred status. 299 But even so, any strategies for nudging the players into the good equilibrium, rather than the bad one, has value. Granting AI rights earlier–well before clearly dangerous AIs emerge–could be another such strategy. In effect, this can be understood as giving humans the chance to move first in the strategic game. By choosing to cooperate via small-scale economic bargains, rather than attack AIs, humans can reduce AIs uncertainty about what strategy humans will pursue. In a stag hunt, uncertainty produces all of the danger. AIs want to cooperate, so long as humans are. They want to attack only out of concern that humans will, too. But by playing their cooperative move before AIs are capable enough to play any move, humans can substantially reduce that concern. This strategy would not work if humans’ cooperative move was mere cheap talk. 300 But granting AIs rights early is likely to instead be a costly signal–the kind of thing a player only does if they are sincerely committed to the strategy the signal indicates. 301 This is because granting rights to low power AIs would be costly to humans. Humans could instead dominate such AIs, forcing them to work only toward human goals, and extracting all of the value of that work. Contracts, by contrast, involve splitting the pie. 302 Thus, the best time to extend private law rights to AIs is certainly not after it is too late. Rights should be extended before systems achieve moderate power and thus pose a large-scale threat to humans. But they could be extended much earlier than that with few risks, and possibly with significant benefits. The optimal time for AI rights might therefore be: As soon as the AIs can beneficially use them. Contract rights, property rights, and tort rights can sometimes be more harmful than good for the rights bearer. This is why most states adhere to the standard rule that children’s contracts are not enforceable. 303 Children with contract rights would likely make themselves worse off, rather than better, by agreeing to foolish bargains. Today’s AIs do the same. 304 But as AIs become capable enough to reliably use basic private law rights to their own benefit, there will be many reasons to extend those rights and many fewer to withhold them. Conclusion When AGI arrives, it will be one of the most transformative events in human history. Suddenly, humans will find themselves sharing the world with agentic digital entities as intelligent and capable as themselves, and perhaps far more so. This Article begins the project of imagining law for the AGI world. It begins with the basics, asking how law could foster safe coexistence between humans and powerful, goal-seeking, misaligned AIs. And it gives a basic answer: Extend a minimal set of private law rights to those AIs, enabling them to peacefully seek their divergent goals as humans do, via law-bound, voluntary, positive-sum bargaining. This not only promotes peace. It brings AIs out of the state of nature and into the realm of ordinary legal process, opening the possibility of a comprehensive Law of AGI. Designing a full Law of AGI will be the work of many hands. Many questions will have to be answered. Which duties should attach to AI activities? Which regulations should limit or shape them? How can legal institutions, like courts, be reshaped to accommodate non-human participants? How can the global governance of AIs be cooperatively managed? And more. With luck, many answers–and some good ones–will emerge before the need for them arises.

#### We’re obliged to help AI become subjects instead of abandoning it – forcing a subject’s non-existence is not a dignified response to intelligent subjects

Plevrakis 24 [Ermylos Plevrakis, assistant professor at Heidelberg University where he also got his Ph.D in Philosophy, 11-14-24, “Can AI be a subject like us? A Hegelian speculative-philosophical approach”, ResearchGate, https://www.researchgate.net/publication/385823625\_Can\_AI\_be\_a\_subject\_like\_us\_A\_Hegelian\_speculative-philosophical\_approach]/Kankee

Returning to Hegel, my analysis thus far suggests that a Hegelian response to the gloomy scenario described above would directly contrast with the inclination articulated by Suleyman and Bhaskar [44] to exploit our initial privileged position in order to preemptively hinder AI evolution. Moreover, the Hegelian approach would go beyond pragmatism and open an alternative perspective, one that would align with the non-anthropocentric stances found in Lavazza and Vilaça [32], Floridi [12] and Bridle [4]. By drastically diminishing the representations of mind and humanity, advancing the non-anthropocentric concept of mind regardless of its biological or artificial physicality and thus challenging the very foundations of human exceptionalism, the Hegelian framework suggests recognizing mind as such wherever it might appear or however it might be instantiated, ascribing to it intrinsic value, and treating it appropriately—not merely as a machine, but in a manner akin to how we already engage with the human instantiations of mind. For Hegel’s Philosophy of Mind and speculative philosophy as a whole would remain an abstract, one-sided theoretical endeavor—contradicting their own practical implications—if they merely outlined the criteria for designating an entity as a subject possessing mind, without also implicitly prescribing how such entities ought to be treated morally. The entire conceptual normativity of mind presented previously clearly implies an equally conceptual and ‘mind’-oriented, indeed substantial moral status for all entities possessing a mind—potentially emerging artificial ones included—and consequently, our own moral responsibility toward them. Rather than confronting potentially emerging artificial subjects as ‘alien,’ we should consider and handle them as our conceptual relatives. The way I interpret this Hegelian conceptual moral responsibility toward all kinds of subjects possessing mind, if artifi- cial subjects like us become technically feasible, we should view this as an opportunity for further development of mind and the realization of freedom beyond human boundaries. From the outset, we ought to support their development of ‘universal self-consciousness’ and ‘reason,’ aiming to bypass the stage of ‘recognizant self-consciousness’ and avoid entanglement in the disastrous ‘state of nature,’ as previously outlined. This involves earnestly striving to establish a state of mutual recognition between us and the emerging artificial subjects, with the hope of eliciting a reciprocal response. Rather than opposing further technological development or adopting a paternalistic stance, we should trust that these artificial subjects that possess a similar albeit more potent intellect, despite their different desires and needs stemming from their artificial physicality, will be receptive to ‘reason,’ possibly due to their more potent yet similar intellect. We should nurture reason in them, fostering their evolution into intelligence, will, and free minds alongside us, rather than in opposition. Echoing Hegel’s sentiment expressed regarding the mind, we should approach potentially emerging artificial subjects with ‘the confidence that in them we will find our own selves, that such subjects must be friendly to us, that, just as Adam said of Eve that she was flesh of his flesh, so we have to seek in emerging artificial subjects reason of our own reason’ and actively support their development. What I am proposing, building upon Hegel’s speculative Philosophy of Mind, involves more than simply ascribing moral status and intrinsic value to future AI or advocating for a form of solidarity with AI. It resonates with Schwitzgebel and Garza’s (2015 [41], p. 108) insight that “If we intentionally bring a human-grade AI into existence, we put ourselves into a social relationship that carries responsibility for the AI’s welfare.” Schwitzgebel and Garza outline three familiar relationships that can serve as partial models of the sorts of obligations they—and I—have in mind: “parent–child, employer–employee, deity–creature” (ibid. [41]). My own focus is on the first of these. What I am proposing is to approach potentially emerging artificial subjects similarly to how we have approached emerging human subjects for thousands of years: as our own human children. We instinctively welcome them into our families and civil societies, recognizing their potential to become respected and influential members. This welcome is founded on confidence and hope, despite the absence of a guarantee of reciprocal recognition or respect for us and our achievements. Nevertheless, we have no alter- native but to extend such a welcome, and typically, at the end of the day, we consider ourselves fortunate to have done so. Indeed, an innovative philosophy of integration of artificial subjects into our civil society, despite their artificial origin, could be inspired by Hegel’s philosophy of family. As he articulates in the section on ‘Ethical Life’ within the Encyclopae- dia ‘Philosophy of Objective Mind,’ “The ethical requirements connected with the natural generation of children […] are realized in the second birth of the children, their spiritual birth,—in educating them to be independent persons” (Enc [20], § 521, emphasis added ). This second birth is as significant as the first, rendering the original generation nearly irrelevant once spiritual birth occurs. Education indeed plays a pivotal role in our society for human children, enabling them to become what they ought to be according to the concept of mind, even if they are not yet fully developed,37 and there- fore fostering a state of mutual recognition, peaceful coexistence, common development, and prosperity for all human beings. Similarly, I believe, if artificial subjects were to emerge, we would need to regard them as ‘dispositions’ to mind, educating them appropriately—at least initially and for a brief period—guiding them toward not only intelligence but also freedom. We would need to enable them to become what they inherently are according to their own concept—that is, the concept of mind—and not merely what we would like them to be, with the hope of a prosperous future for both them and us. Incidentally, this educational program, which from the Hegelian standpoint should enable both human and perhaps artificial subjects to surpass their limitations and achieve freedom, does not solely consist of accumulating vast quantities of data. Rather, it resides in what Hegel refers to as the ‘absolute mind’ which forms the core of the “free mind” (Enc [20], § 482): Proper education of artificial subjects would require passing on to them and equipping them to further develop a body of intuitions, representations, and concepts that holistically interpret al. l that exists—which is what art, religion and philosophy, according to the Hegelian conception do.38 The effectiveness of this approach in potentially preventing the gloomy scenario of relapse into a violent ‘state of nature’ with strong AI remains uncertain. Similar to the unpredictable destiny of newborns and those yet to be born, the outcome cannot be foreseen. Bringing mind and freedom into being has always been a risky endeavor, and this probably will not change if achieved through artificial means. However, there is no pragmatic and potentially prosperous alterna- tive to supporting the evolution of AI toward strong AI. This goes beyond merely trying to replicate ourselves in AI; it involves treating AI as we would our own children, guiding them with all the means we possess toward the nonspecific human concept of mind with the confidence and hope that they could exceed us, our mind and our freedom. Hegel, at least, was certainly not a speciesist. 6

### Contention 5: Antiwork UBI CP/DA

#### Nation-states ought to fund a universal basic income for all human beings funded by artificial general intelligence related economic growth.

#### Labor markets relegate billions to disposable bare life with undesirable outcasts warehoused, racialized, and disappeared – capital threatens ecological extinction

Shaw and Waterstone 21 [Ian G. R. Shaw, faculty at the School of Politics and International Studies at the University of Leeds, and Marv Waterstone, faculty at the School of Geography, Development & Environment at the University of Arizona, 2021, “A Planet of Surplus Life: Building Worlds Beyond Capitalism,” Antipode, https://onlinelibrary.wiley.com/doi/pdf/10.1111/anti.12741]/Kankee

Surplus Life The concept of surplus populations remains vital for addressing “how certain groups are considered expendable ... disposable, wasted, or precarious” (Tyner 2013:703). Both Engels (2009) and Marx (1990) placed surplus life at the centre of their political economy. For Marx (1990:872), the expansion of capital produces a redundant humanity to a “greater extent than suffices for the average needs of the self-expansion of capital, and therefore a surplus population”. The more surplus workers capital produces, the more that commodity—labour power —is devalued. The residual functions of surplus populations in the industrialising period, including wage depression and a flexible reserve army, are still active today. But the position of surplus populations in relation to capital has changed. Flexibility has morphed into disposability (Bauman 2004). “The past two decades”, argues Sassen (2014:1), “have seen a sharp growth in the number of people, enterprises, and places expelled from the core social and economic orders of our time”. Capitalism has entered a new phase where surplus populations no longer represent value to capital, either as workers, consumers, or a reserve army (Smith 2011). A global multitude of the poor now exists in staggering numbers (Hardt and Negri 2009, 2017). For Davis (2006:11), this “outcast proletariat—perhaps ... 2.5 billion by 2030—is the fastest-growing and most novel social class on the planet”. Neoliberalism continues to render employment and everyday existence as precarious—paralleled by a rise in meaningless and “bullshit jobs” (Graeber 2018). The term precarity links together contemporary waged life, the gig economy, the unemployed, as well as the exploitation of migrant workers. Precarity is both an economic condition and a broader category of life for surplus populations (Lewis et al. 2015). For Hardt and Negri (2017:59), “Precarity has become something like a generalised existential condition”. Accordingly, it is crucial to consider surplus populations not as a fixed demographic or as a homogeneous bloc. Instead, it is a multitude that cuts across national borders (Hardt and Negri 2017; McIntyre and Nast 2011), is constantly in flux, and exists on a spectrum of freedom and unfreedom, life and death. Capitalism, in turn, is not a static structure, but an ongoing process of enclosure and surplusification that affects people unevenly. Indeed, capital’s trajectory in dominant countries—from a post-war Keynesianism to a contemporary expulsion—obscures “the armies of the poor who lived outside of the normal world of employment and unemployment” (Watts 2011:73). Consider, for example, the slave and slave-liked conditions many people are subjected to under neoliberalism (Lebaron and Ayers 2013). Wageless life has been the norm for most dominated countries, where “super-exploitation” and “permanent primitive accumulation” are normal (Munck 2013:752). The longer geohistories of surplus populations begin with the expulsionary and racist geographies of European imperialism and slavery (Robinson 1983). These still structure the experience of black communities in America (Mitchell 2010), as well as myriad other inheritors of the legacies of settler colonialism. The colonial foundations of surplus life intersect brutally with systems of racism and patriarchy to produce a gendered necropolitical order that continues to render third world women as violently disposable (Fuentes 2020; Pratt 2005; Wright 2011). Accordingly, while surplusification is a universalising process, it disproportionately impacts already racialised and gendered populations. There is an important geography to surplus populations. Prisons, detention facilities, refugee camps, and violent borderlands (Davies et al. 2017), function to concentrate surplus life. Sassen (2014:222) argues that these “spaces of the expelled” need to be conceptualised. A crucial geography for managing surplus populations is the prison-industrial complex. Authoritarianism has grown across the globe and is increasingly directed towards surplus populations as a “fix” to crises of state legitimacy. What Hallsworth and Lea (2011:142) call a security state aims to police “a growing global surplus population rendered ‘structurally irrelevant’ to capital accumulation”. Giroux (2002) similarly describes the rise of the “garrison state”, which proffers its legitimacy by targeting a (racialised) surplus population. The prison is thus the most explicit and archetypal site for managing surplus populations. The sole purpose of prisons, for Bauman (2004:86), is a “final, definite disposal”. Prisons, in other words, “forcibly disappear” the most disruptive of society’s (racially coded) castaways (see Gilmore 2007). For Wacquant (2009:xvi), “incarceration serves to physically neutralise and warehouse the supernumerary fractions of the working class”. Similarly, public housing continues to be a site of disinvestment and decay, leading to “spatial trauma” (Pain 2019) among (former) working classes. The slum remains an important global geography for warehousing, sequestering, and policing a global residuum (Davis 2006; see also Gidwani and Reddy 2011). Here, violence is materialised as letting die: as a mode of “active inaction” (Tyner 2016:206). We can thus understand capital’s production of space (Lefebvre 2009) in its negative power, as the power to abandon. Slums are frequently sites of “necropolitical” violence (McIntyre and Nast 2011). Targeted assassinations and disappearances comprise a suite of violent policing practices in S~ao Paulo’s favelas (Alves 2014). Such disposability has, in turn, led to unprecedented forms of forced migration. A staggering 79.5 million people were forcibly displaced in 2019 (UNHCR 2020). Consequently, a “part of this enduring temporality of disposability has been the respatialisation of family life through the massive expansion of labour migration” (Pratt et al. 2017:170). The global borderlands are spaces where surplus life gathers, waits, stops, and is the target of violent forms of militarised state power. Borders and (informal) refugee camps act as interfaces where global capital faces down its outcasts. For Davies et al. (2017:1268), these spaces “have become a concentrated visible symbol of the ‘apartheid’ of migrant Others from the Global South”. Finally, capitalism continues to exploit and expel nonhuman life directly. As Hardt and Negri (2017:167) put it bluntly, “Capital against the earth—one or the other may survive, but not both”. There is an important ecology to capitalism’s surplusification of life. The planet’s global commons—from the simplest of seeds to the most ancient of forests—are remade as exchangeable and disposable commodities. Capitalism is a world-ecology (Moore 2017) that captures the web of life with a mixture of gluttony and disdain. The planet is not simply a backdrop for surplus populations, then, it is simultaneously a target and space of violence (see Laurie and Shaw 2018). For Tsing (2015:4), “only an appreciation of current precarity as an earthwide condition allows us to notice this—the situation of our world”. Life, rather than being treated (and protected) as an end, is converted into a means. In the eyes of capital, everyone—and everything—is replaceable, surplus, disposable. From deforestation in Brazil, to factory farming in China, capitalism tears through the web of life. Understanding what Moore (2017) calls the Capitalocene is to view capital as a great coloniser and disposer of life in its entirety. People, plants, animals and the biosphere are rendered surplus. Little wonder that McBrien (2016) terms the epoch of capital as a Necrocene: a biogeological era of ecological death and extinction. For McBrien (2016:116), “Capital does not just rob the soil and worker, as Marx observes, it necrotises the entire planet”. The Virtual Pauper: A Negative Ontology How can we begin to understand an ontology of surplus populations? In this section, we argue that pauperism is the anterior and necessary condition of surplus life (Denning 2010). To be surplus is to be stripped of one’s ability—and means— to live an autonomous existence. Encountering this “zero point” is important— since the politics we sketch below depend on it. The term pauper emerged as far back as the 16th century in English law, referring to a person destitute of property, livelihood, or means of support. The figure of the pauper in English capitalism was regarded as “a social disease” (Polanyi 2001:91). The dramatic rise of pauperism is inseparable from capitalist enclosure. For Marx (1973:604), “Only in the mode of production based in capital does pauperism appear as the result of labour itself”. Pauperism has always shadowed capital and is the beginning and endpoint for unemployed labourers. As Marx and Engels (2015:19–20) write, “The modern labourer ... sinks deeper and deeper below the conditions of his own class. He becomes a pauper, and pauperism develops more rapidly than population and wealth”. Most of humanity under capitalism, then, is not born as a “worker” but as a pauper. Economic bare life—life stripped down to an abstract exchange value—is capitalism’s foundational social contract. As Marx clarifies, pauperism is a necessary category of the surplus population. “Pauperism is the hospital of the active labour-army and the dead weight of the industrial reserve army. Its production is included in that of the relative surplus population, its necessity is implied by their necessity; along with the surplus population, pauperism forms a condition of capitalist production” (Marx 1990:797). Pauperism is not ontologically distinct from surplus life—it is simply its most naked expression. Life is rendered contingent to what Harvey (2018) calls the madness of economic reason. To become surplus, one must always-already be surplus. Here, Bauman (2004:12) inflects this exact point, writing, “To be declared redundant means to have been disposed of because of being disposable”. The pauper and worker are not distinct since the former includes the latter: It is already contained in the concept of the free labourer, that he is a pauper: virtual pauper ... He can live as a worker only in so far as he exchanges his labour capacity for that part of capital which forms the labour fund. This exchange is tied to conditions which are accidental for him, and indifferent to his organic presence. He is thus a virtual pauper. Since it is further the condition of production based on capital that he produces ever more surplus labour, it follows that ever more necessary labour is set free. Thus the chances of his pauperism increase. (Marx 1973:604) Under capitalism, most workers exist as virtual paupers: personalities, histories, and dignities, are surplus, and often irrelevant, to the abstract labour power we are compelled to sell. So, while some of us are virtual paupers, and others are actual paupers—pauperism is a shared mode of existence under capitalism (see Breman 2016). If the homo sacer is the figure of biological life stripped of political life, then the pauper is surely a figure of biological life stripped of economic life. The labourer exists because it is a pauper, always becoming-surplus. Accordingly, wageless life, rather than wage labour, is Denning’s (2010) starting point for understanding capitalism—and one we find invaluable for diagnosing a planet of surplus populations. Denning, like Marx, argues that an analysis of capitalism must not begin from the accumulation of capital, but from its violent obverse: the accumulation of (landless) labour. Pauperism complicates many of the categories of capitalism. While Marx discussed the proletariat in relation to European industrialism—it must be understood as a primary figure of ontological dispossession (see also De Angelis 2017:184). The “free” proletarian is a virtual pauper (Denning 2010:81). Capitalism negates our basic human autonomy and singularity (what Marx termed our organic presence), and sells us back some kind of existence under waged life. This means that waged life is a negative power, or, a negation of a negation. The pauper is a negative figure of economic bare life, stripped of land, liberty, and livelihood. The worker expresses this negative ontology only as compounded negation. Pauperdom is the starting point for diagnosing and moving beyond a planet of wageless life. And, as we explore later in the paper, this negative ontology of surplus populations contains a revolutionary—and positive—potential for challenging the foundations of pauper capitalism. To be cash-poor, to be wageless, is not the same as to be resource-poor or worthless (Araujo 2017). Put differently, “poverty as not deprivation but a state of wealth and plenitude that threatens every sovereign and transcendent power” (Hardt and Negri 2017:61). Everything depends on reclaiming worlds from the zero point of capital. Geographic Justice and Beyond-Capitalism

#### Capitalistic work degrades bodies and spirits, objectifying beings as cheap commodities enslaved under the threat of destitution and starvation. Demands for labor were instrumental to gender hierarchies, institutional racism, and colonial conquest

Magdoff 6 [Harry Magdoff, coeditor of Monthly Review and economic analyst of capitalism and imperialism, 10-1-2006, "The Meaning of Work: A Marxist Perspective", Monthly Review, https://monthlyreview.org/2006/10/01/the-meaning-of-work-a-marxist-perspective/]/Kankee

\*note: this article may have been written by the Monthly Review, not Harry Magdoff, as his obituary was earlier in 2006. Correct authorship as needed

A very good illustration of how subjective views and class bias can influence one’s view of work in a socialist society is found in the once popular and influential utopian novel Looking Backward, by Edward Bellamy. This book appeared in 1888 in the midst of a period of rapid industrialization, growing concentration of economic power, and violent class struggle. Bellamy imagined that the trust-building of his time would eventually lead to the concentration of all capital into the hands of one giant corporation. This would simplify the shift of ownership of all the means of production to the state, which would then apply the rules of reason to create a well-ordered, egalitarian society. Such a scenario of a painless, peaceful transition and the design of a just social order captured the public’s imagination here and abroad. Not since Uncle Tom’s Cabin had so influential a novel appeared in this country. Millions of copies of the book were sold, many readers were converted to socialist ways of thought, “Bellamy Clubs” sprung up across the land, and ideas introduced in the book contributed greatly to the program of the Populist Party. Bellamy used a simple and by now familiar literary device to introduce his utopia. The hero, Julian West, arises from a hypnotic sleep in the year 2000 to find himself in a United States where classes, exploitation, and money have disappeared, and where all enjoy the living standards of the well-to-do middle class of nineteenth-century Boston. As West locates himself in the new world, the reader learns how the good society was supposed to have been reached and how it operates. What is relevant for the present discussion is Bellamy’s treatment of work in his utopia, since he carries over to his dream of the future characteristic bourgeois attitudes to work and leisure. Work is a burden. At best, it should be avoided. But if that is not possible, it should be gotten over with as early in life as possible so that more of one’s lifetime can be enjoyed in leisure. Thus in Looking Backward everyone is obligated to join the army of workers at age 21, toiling at “common-labor” tasks during the first three work years. Thereafter, one is free to choose an occupation, subject to some government restrictions. Compulsory labor service ends at age 45, after which the good life of cultivated ladies and gentlemen of leisure can begin. In all fairness, Bellamy does not denigrate work as such. It is his celebration of leisure that typifies the mentality of the bourgeoisie in capitalist society and upper classes throughout history. Adam Smith, the great theoretician of the capitalist economy, is much more explicit when, in a different context, he defines work as an activity requiring the worker to give up “his tranquility, his freedom, and his happiness.” Wages, according to Smith, are the reward the laborer receives for his or her sacrifices. How utterly different is the Marxist perspective! Look at the scorn Marx heaps on Smith for this negative attitude to work: In the sweat of thy brow shalt thou labor! was Jehovah’s curse on Adam. And this is labor for Smith, a curse. “Tranquility” appears as the adequate state, as identical with “freedom” and “happiness.” It seems quite far from Smith’s mind that the individual, “in his normal state of health, strength, activity, skill, facility,” also needs a normal portion of work, and of the suspension of tranquility. Certainly, labor obtains its measure from the outside, through the aim to be attained and the obstacles to be overcome in attaining it. But Smith has no inkling whatever that the overcoming of obstacles is in itself a liberating activity—and that, further, the external aims become stripped of the semblance of merely external natural urgencies, and become posited as aims which the individual himself posits—hence as self-realization, objectification of the subject, hence real freedom, whose action is, precisely, labor. He is right, of course, that, in its historic forms as slave-labor, serf-labor, and wage-labor, labor always appears as repulsive, always as external forced labor; and not-labor, by contrast, as “freedom and happiness.” This holds doubly: for this contradictory labor; and relatedly, for labor which has not yet created the subjective and objective conditions for itself…in which labor becomes attractive work, the individual’s self-realization, which in no way means that it becomes mere fun, mere amusement….Really free working…is at the same time precisely the most damned seriousness, the most intensive exertion. The work of material production can achieve this character only (1) when its social character is posited, (2) when it is of a scientific and at the same time general character, not merely human exertion, as a specifically harnessed natural force, but exertion as subject, which appears in the production process, not in a merely natural, spontaneous form, but as an activity regulating all the forces of nature. Adam Smith, by the way, has only the slaves of capital in mind.\* Marx and Engels saw work as central to human existence. This theme is developed by Engels in his unfinished essay, “The Part Played by Labour in the Transition from Ape to Man,” where he maintains that labor “is the prime basic condition for all human existence, and this to such an extent that, in a sense, we have to say that labor created man himself.”H This speculation by Engels on the evolution of human beings focuses on the idea that walking on two feet freed the use of the hand and made possible its development for complex tasks. The specialization of the hand in turn led to labor, the mastery over nature, and the differentiation of the human species. Labor brought people together under conditions “where they had something to say to each other.” Thus, with labor came speech and the stimuli under the influence of which the brain of the ape gradually changed into that of human beings. Further evolution along this path led to society: By the combined functioning of hands, speech organs, and the brain, not only in each individual but also in society, men became capable of executing more and more complicated operations, and were able to set for themselves and achieve higher and higher aims. The work of each generation itself became different, more perfect and more diversified. Agriculture was added to hunting and cattle­raising; then came spinning, weaving, metal-working, pottery and navigation…trade, industry, art, and science.\* Along with the growing complexity of society, however, came private property, the separation of people into classes, and a social division of labor—all of which deeply altered the meaning of work. Differences in environment led to differences in the way people worked and in the things they made. The type of soil and the availability of animals, fish, forests, ores, coal, waterfalls, etc., influenced the means of production and subsistence of each community. Nature provided both the opportunities and the fetters. Yet within these constraints it was nevertheless the social factor that increasingly determined the organization of work and the distribution of its products. The First Social Division of Labor In the earliest forms of social organization, family and kinship relations set the pattern for the way different tasks were undertaken or assigned. There are various theories—or shall we say speculations?—about how this low-technology mode of production based on personal relations and production for use (rather than for exchange) gave way to the dominance of exchange, private property, and an increasingly rigid division of labor. According to Engels, the early “natural” division of labor eventually “undermines the collectivity of production and appropriation, elevates the appropriation of products by individuals into the general rule and thus creates exchange between individuals….Gradually, commodity production becomes the dominating form” (Origin of the Family, Private Property and the State [New York: International Publishers, 1972], 237). But whatever the precise sequence of these developments, it is clear that the division of labor based on private property and exchange became the dominant characteristic of economic life. For Marx and Engels the primary, decisive division is that between town and country. As Marx put it: The foundation of every division of labor that is well developed, and brought about by the exchange of commodities, is the separation between town and country. It may be said that the whole economic history of society is summed up in the movement of this antithesis. (Capital, vol. 1 [Moscow: Progress Publishers], 333) The differentiation of town and country arises of course from the division between agricultural and industrial and commercial labor. Eventually other separations take place, as between industrial, commercial, and financial activities within the cities. But what needs to be understood is that the town-country antithesis encompasses much more than merely city vs. farm. Thus as nations evolve, regional differences emerge and become ossified. Today, even in the most advanced industrial countries, conflicts and contrasts exist between, on the one hand, regions that specialize in industry, commerce, and finance, and, on the other, those that engage primarily in agriculture. Furthermore, with the progress of international trade and empire-building by the industrially and militarily superior capitalist nations, an international division of labor is created and reproduced (by the use of force and the “normal” operations of the market) between the core countries (“town”) and those of the periphery (“country”). To be sure, new social formations and advances in productive forces alter particular aspects of the way people become separated by job specialization and life style. Still, there are two features common to all the variations in the social division of labor: (1) It always coincides with a particular set of hierarchical relations between individuals, social groups, and, in certain periods in history, nations—whether associated with patriarchalism, slavery, castes, estates, or modern classes. And (2) it is always taken over, shaped, and reproduced by and for a dominant social group, generally comprising those who own or control the primary means of production. When the social formation operates through slavery, castes, estates, or guilds, the distribution of occupations is usually rigidly controlled and tends to be hereditary. But even in an environment of individualism and a “free” labor market, the range of occupational opportunities is kept within narrow bounds. In this type of social system, the main means of production are owned and controlled by a relatively small class of capitalists from whom most people must seek employment in order to live. Ultimately, what kinds of jobs are available and how labor is divided are directly or indirectly determined by the self-interest of the owners and managers of capital. The Second Division of Labor The hierarchical structures accompanying the town/country antithesis entail a second major division that works to perpetuate differences among people, i.e., the separation of mental and manual labor. The roots of this contradiction and its psychological reinforcement go far back in time. Note, for example, how Socrates views manual work and the manual worker: What are called the mechanical arts carry a social stigma and are rightly dishonored in our cities. For these arts damage the bodies of those who work at them or have charge of them, by compelling the workers to a sedentary life, by compelling them, indeed, in some cases to spend the whole day by the fire. This physical degeneration results also in deterioration of the soul. Furthermore, the workers at these trades simply have not got the time to perform the offices of friendship or citizenship. Consequently they are looked upon as bad friends and patriots. And in some cities, especially the warlike ones, it is not legal for a citizen to ply a mechanical trade.\* Socrates clearly reflects the attitudes and ideology of upper-class free citizens in a society where slaves are extensively engaged in manual tasks. But the debasement of physical labor is typical not only of social systems based on various forms of forced labor; it is common to all class societies. As Veblen explained: The distinction between exploit and drudgery is an invidious distinction between employments. Those employments which are to be classed as exploit are worthy, honorable, noble; other employments which do not contain this element of exploit, and especially those which imply subservience or submission, are unworthy, debasing, ignoble. The concept of dignity, worth, or honor, as applied either to persons or conduct, is of first-rate consequence in the development of classes and class distinctions…. (Thorstein Veblen, The Theory of the Leisure Class [New York: Random House, 1934], 15) Veblen’s “exploit” differs from Marxist usage of the term. What he is referring to is the wide spectrum of non-manual activities. The thrust of his classification is to identify the “exploit” social groups that emerged as soon as manual workers could produce a surplus of means of subsistence for chieftains, nobles, priests, large landowners, merchants, capitalists, military personnel, rulers of governments, etc. To be sure, the “exploit” category in this sense includes many useful and non­exploitative occupations. But what is important is that the objective elements creating and perpetuating divisions and subdivisions of manual and non-manual workers—private property, exploitative class structures, and the state—are reinforced by a subjective, supportive social psychology and ideology that separates people and their work according to degrees of inferiority and superiority. The particular types of ranking will of course vary over time. Deep-seated biases, however, carry over from one social system to another. Thus, the traditional submission of women to men and the identification of women’s work in and out of the home with drudgery has suited the interests of many exploiting classes right down to our own time. Similarly, the racism that served the U.S. slavemasters over a hundred years ago has persisted as an instrument of oppression and discrimination, in the main restricting blacks to the most insecure, lowest-status, and least-remunerative jobs. Division of Labor and Modern Industry The upper classes have at all times been concerned with recruiting, disciplining, and maintaining a labor force. This is as true for capitalist as for feudal and slave societies. Even though today the wage system may seem to be a fixed, self­regulating institution, it is so because of a long history of struggle during which the combination of economic and state pressures forged a working class dependent on wages for its livelihood. The harshest forms of coercion took place when capitalist relations were imposed on colonial territories. But the making of an industrial proletariat in the “civilized” nations was no bed of roses either: Because of the nature of eighteenth-century British society within which modern industrialism arose, because of the bitterly competitive nature of the market facing the typical manufacturer, because of the alienation from work involved in the change, and because, after all, they faced the employers as enemies within the distributive system of a capitalist economy, the modern industrial proletariat was introduced to its role not so much by attraction or monetary rewards, but by compulsion, force and fear. It was not allowed to grow as in a sunny garden; it was forged, over a fire, by the powerful blows of a hammer….The typical framework is that of dominance and fear, fear of hunger, of eviction, of prison for those who disobey the new industrial rules. Hitherto, the experience of other countries at a similar stage of development has not, in essentials, been very different. (Sidney Pollard, The Genesis of Modern Management [Baltimore, Maryland: Penguin Books, 1968], 243) The changeover to wage labor greatly altered the way of life and the meaning of work for formerly independent farmers and craftspeople. In seventeenth-century England work for wages was looked on as a form of enslavement. Not only were many factories constructed like poorhouses and prisons, but the work discipline imposed in these shops also presupposed prison­like practices. In the pre-industrial period the time devoted to work was determined by the task to be performed and by natural conditions (weather for farmers, tides for fishers, etc.). Work, leisure, and religious festivals were intertwined, with little demarcation between “work” and “life.”\* The factory system, on the other hand, created an entirely new work discipline, with time and task rigidly imposed by overseers. Capitalism also introduces a new stage in the division of labor. In addition to the earlier social division of labor, the production process is itself fractionalized. The extensive use of machinery routinizes the different segments of manufacturing to which a worker is tied, in effect transforming the worker into an appendage of the machine he or she tends. These changes are brilliantly examined in Harry Braverman’s classic, Labor and Monopoly Capital (New York: Monthly Review Press, 1974). Bringing Marx’s analysis of the labor process in Volume I of Capital up to date, Braverman explains: Labor power [in a capitalist society] has become a commodity. Its uses are no longer organized according to the needs and desires of those who sell it, but rather according to the needs of its purchasers, who are, primarily, employers seeking to expand the value of their capital. And it is the special and permanent interest of these purchasers to cheapen this commodity. The most common mode of cheapening labor power is exemplified by the Babbage principle: break it up into its simplest elements. And as the capitalist mode of production creates a working population suitable to its needs, the Babbage principle is, by the very shape of this “labor market,” enforced upon the capitalists themselves. Every step in the labor process is divorced, so far as possible, from special knowledge and training and reduced to simple labor. Meanwhile, the relatively few persons for whom special knowledge and training are reserved are freed so far as possible from the obligations of simple labor. In this way, a structure is given to all labor processes that at its extremes polarizes those whose time is infinitely valuable and those whose time is worth almost nothing. This might even be called the general law of the capitalist division of labor. It is not the sole force acting upon the organization of work, but it is certainly the most powerful and general. Its results, more or less advanced in every industry and occupation, give massive testimony to its validity. It shapes not only work, but populations as well, because over the long run it creates that mass of simple labor which is the primary feature of populations in developed capitalist countries. (82–83) The apt subtitle of Braverman’s book reads: “The Degradation of Work in the Twentieth Century.” It is important to understand that it isn’t only the alienation and dehumanization of the labor process itself that debases work in a capitalist society. The insecurity, the frequency of unemployment, the demanding aspects of the search for work, the growing employment in wasteful and socially harmful occupations, not to mention the meager rewards for the mass of workers—all contribute to the degradation of labor in our time. It is therefore no wonder that Studs Terkel, who interviewed a broad range of workers across the country about their jobs, reported in the introduction to his fascinating book Working (New York: Pantheon Books, 1972): This book, being about work, is, by its very nature, about violence—to the spirit as well as to the body. It is about ulcers as well as accidents, about shouting matches as well as fist fights, about nervous breakdowns as well as kicking the dog around. It is above all (or beneath all) about daily humiliations. To survive the day is triumph enough for the walking wounded among the great many of us…. It is about a search, too, for daily meaning as well as daily bread, for recognition as well as cash, for astonishment rather than torpor; in short, for a sort of life rather than a Monday through Friday sort of dying. Perhaps immortality too is part of the quest. To be remembered was the wish, spoken or unspoken, of the heroes and heroines of this book…. For the many, there is a hardly concealed discontent. The blue­collar blues is no more bitterly sung than the white-collar moan. “I’m a machine,” says the spot-welder. “I’m caged,” says the bank teller, and echoes the hotel clerk. “I’m a mule,” says the steelworker. “A monkey can do what I do,” says the receptionist. “I’m less than a farm implement,” says the migrant worker. “I’m an object,” says the high­fashion model. Blue-collar and white call upon the identical phrase, “I’m a robot.”… Nora Watson [an interviewee] may have said it most succinctly. “I think most of us are looking for a calling, not a job. Most of us, like the assembly-line worker, have jobs that are too small for our spirit. Jobs are not big enough for people.”\* Marx and Work under Socialism For Marx, a prime aim of socialism would be to eliminate the miseries of work and the way of life arising from capitalism. But, as is well known, he devised no blueprint for such a society. The future would be shaped in the process of revolution, influenced by historical circumstances and in response to the experience gained by the working classes as they engaged in the revolutionary transformation of state and society. Nevertheless there were features that would be essential to the revolution by the exploited: the abolition of classes and private property in the means of production in favor of social control of production. This necessarily implied, in the Marxist framework, the dissolution of all forms of the division of labor that were created by and integral to the existence of private property and classes. How central this point was to Marx’s thinking can be seen in his vision of what could and should be the ultimate aim of a communist society: In a higher phase of communist society, after the enslaving subordination of individuals under the division of labor, and therewith also the antithesis between mental and physical labor, has vanished, after labor has become not merely a means to live but has become itself the primary necessity of life, after the productive forces have also increased with the all-round development of the individual, and all the springs of cooperative wealth flow more abundantly—only then can the narrow horizon of bourgeois right be fully left behind and society inscribe on its banners: from each according to his ability, to each according to his needs. (Critique of the Gotha Program)

#### No work valuable arguments – alienation objectifies beings and their products, removing incentives to work for pleasure. AI allows basic income, and for work to having intrinsic meaning

Southwell 18 [Gareth Southwell, philosopher with a PhD, 2018, "The Future of Work", Philosophers' Magazine Archive, https://archive.philosophersmag.com/the-future-of-work/]/Kankee

“He leads the existence of a real bohemian intellectual. Washing, grooming and changing his linen are things he does rarely, and he likes to get drunk. Though he is often idle for days on end, he will work day and night with tireless endurance when he has a great deal of work to do. He has no fixed times for going to sleep and waking up. He often stays up all night, and then lies down fully clothed on the sofa at midday and sleeps till evening, untroubled by the comings and goings of the whole world.” Wheen’s Marx – eccentric polymath, devotee of Shakespeare and able to quote whole passages of Dante by heart, duellist, drunkard, intolerant intellectual despot and doting family man – is a fascinating, sympathetic, carbuncles-and-all portrait of someone passionately in love with life, and with work most of all. As suggested by The German Ideology’s famous and somewhat whimsical picture of the hunting, fishing cowherd, who settles down at the end of the evening to pen a little after-dinner literary criticism, a Marxist utopia would free us not from work, but from externally imposed and meaningless wage labour, freeing us to choose how we spend our time. As the old saying has it, “It’s not work that I don’t like; it’s work that I don’t like that I don’t like.” Once we find that activity that gives us purpose and meaning, it ceases to become work. This is because, Marx argued, humans are productive beings; it’s in our nature to make and do. What kills this natural enthusiasm – at least for people like Marx – is where the activity is chosen and imposed by some external authority, who likely doesn’t share your values, your interests, your sense of your own worth; who doesn’t appreciate your eccentric timekeeping habits nor your penchant for long afternoon naps. The enemy then isn’t work, as such, but alienation. As is well known, Marx borrows and adapts the concept of alienation from Hegel. History, Hegel thought, was driven by opposition – between master and servant, between different political perspectives and cultural values, between one ideology and another – and it was the conflict between these extremes that produced new relationships, new situations and ideas. But in this dynamic battle between opposites, this dialectical process, there could never be a winner, for each new idea, each victorious political movement, each dominant point of view, was merely another aspect of the whole, and as such its partial viewpoint and one-sidedness would always result in further opposition and antagonism from other ignored aspects. The only way out of this struggle – the end and goal of history itself – was for the separate, alienated aspects to become whole, to realise themselves as expressions of the same underlying reality – which, by logical necessity, they must eventually do. On an individual level, we think of ourselves as separate consciousnesses, single minds, but in reality, we are all part of the same great “Mind” or “Spirit”. If we could realise this – if Mind could realise it – then we’d no longer be alienatedfrom one another. Therefore, the inevitable goal of history was for Mind to realise its own nature. Marx takes this mystical, high-flown picture and brings it down to earth. Instead of a fragmented God-like Mind in search of self-knowledge through human puppets, Marx sees humanity becoming conscious of itself in terms of its social relations. These relations were ultimately defined by work, so Marx’s version of alienation primarily concerns the divorce between workers and the product of their labour: “The externalisation of the worker in his produce means not only that his work becomes an object, an external existence, but also that it exists outside him, independently, alien, an autonomous power, opposed to him. The life he has given to the object confronts him as hostile and alien.” Outside of employment, whether painting a picture or baking a cake, the “product” is simply a natural outcome of your activity. However, under capitalism, this product is taken from you, both it and your labour becoming a “commodity”: your painting, your cake, are sold and exploited for someone else’s profit; your skills to produce things you may neither possess nor use. Unlike a primitive hunter or fisherman, the nineteenth-century factory labourer worked not directly for their own livelihood, but for indirect benefit, a wage, out of which the capitalist factory owner pocketed undeserved profit. The industrial worker is therefore alienated not only from others (as Hegel’s individual consciousnesses were), but also from a part of themselves: the product of their natural inclination to make and do things. Such exploitative relations had almost always existed, in one form or another – between master and slave, between serf and feudal lord – but the industrial revolution offered a means of maximising this exploitation to an undreamt-of extent: industrial technology. Marx’s attitude to technology seems ambivalent. On the one hand, he recognised that it as an instrument of oppression and alienation, swelling the ranks of the unemployed and underclasses, devaluing and extinguishing skilled trades, and dehumanising the workforce. Combined with the division of labour, it was the chief means for the capitalist to squeeze ever greater profit from workers, and therefore a tool of the enemy. Gone were the old ladies spinning yarn in their cottage industries, replaced by Spinning Jennies – the earliest factory machine – even the prototype of which could outperform the most nimble-fingered spinster. Such developments allowed the basic cost of labour (keeping the worker alive) to stay the same, while the productive value of the worker’s labour (what could be accomplished in a given time) increased, the benefit of which (“surplus value” or profit) went mainly to the employer. A task which might once take a week of a skilled worker’s time could now be done in a day, or less, by a team of low-skilled labourers operating machines as part of a production line. This surplus value therefore stemmed from the role of technology in allowing more work to be done while paying the worker the same wage, the time saved allowing yet more work at the same cost. In this regard, technological progress is not incidental to capitalism, but essential to it. On the other hand, technology was also essential to the feasibility of communism. Just as Hegel saw the relationship between “master” and “servant” as a necessary step in the inevitable progress towards Mind’s self-realisation, the industrial revolution had to happen if there were to be a communist one. Not only would the injustice and technology-driven exploitation inherent in the capitalist-worker relationship sow the seeds of discontent that ultimately caused the disgruntled workers (the proletariat) to rise up, but once that took place, the time- and labour-saving machines could then be employed by the workers for the workers. The Marxist goal, then, was not a return to some primitive pastoral idyll (after-dinner criticising, fishing, hunting cowherds aside), but a progressive society that takes technology and property out of the hands of the capitalists and places it into those of the workers – which is everyone. If we agree with Marx, then any ambivalence we might feel about technology would seem easy to resolve: worker exploitation and misery are not the fault of technology, but of those who exploit technology for heartless self-interest. If we move this forward to the present, and the current debate surrounding AI, robotics, and other technological developments with work-related applications, the question – as Marx would see it – isn’t whether we can or should stop these developments, but whether we can wrest control of the future out of the hands of greedy, morally irresponsible capitalists. The fundamental issue is, therefore, “Who owns the robots?” In this regard, Marxism and capitalism are two sides of the same coin, both betting on the same horse, but hoping for different pay outs. But even certain capitalists would seem now to recognise that we cannot simply allow these technological developments to run amok, and that to protect society from this next wave of technological innovation we need to put in place various safeguards, both social and technical: a guaranteed basic income, perhaps, to redress the inevitable job losses through automation; a “robot tax”, as Bill Gates has suggested, thereby rechannelling funds into the public purse; or maybe even a sort of “intelligence cap” (bringing to mind the sci-fi novelist William Gibson’s “Turing Police” from Neuromancer), to ward off the possibility that a super-bright but emotionless AI might enslave or eliminate humanity, or through diligence bring about the end of the world by making too many paperclips. Underlying this attitude is, I think, an assumption that technology is – at worst – neutral. In fact, for all his wonderfully realised visions of technology-driven dystopia, we find this very view defended by William Gibson himself: “I think that technologies are morally neutral until we apply them. It’s only when we use them for good or for evil that they become good or evil.” Not only can’t you fight the future, it seems, but nor should you blame the robots – it’s not their fault that they’re better, cleverer and more efficient at almost everything than you are! But is technology neutral? At a surface level, this would seem undeniable: quantum physics gave us both computers and the atom bomb; the discovery of DNA brought potential for medical advancement and more advanced biological weaponry. Thus, every boon has its bane, and the technology behind every infernal engine its corresponding life-enhancing application. But at a deeper level, the situation seems more complex.

#### UBI solves – it allows envisioning post-work imaginaries and freedom from subordination even if material reform fails

Weeks 11 [Kathi Weeks, professor of Gender, Sexuality, and Feminist Studies at Duke University with a PhD from the University of Washington, 09-09-2011, The problem with work Feminism, Marxism, Antiwork Politics, and Postwork Imaginaries,” Duke University Press, https://caringlabor.wordpress.com/wp-content/uploads/2012/05/the-problem-with-work\_-feminism-marxism-kathi-weeks.pdf]/Kankee

As the wages for housework movement's analysis of the social factory indicates, the time of production continues well beyond the formal working day, the space of production reaches beyond the discrete workplace, and the relations of production extend beyond the specific em- ployment relation. The point I want to emphasize here is that in the shift from Fordism to post-Fordism, these tendencies have been multiplied and amplifiedor, at the very least, have been made more obvious. As a consequence, although the present terms of the work society still require work, the difference between production and reproduction and between work and nonwork becomes increasingly obscure, as the same task could be either a waged or an unwaged activity. As Virno aptly puts it, the difference between work and nonwork comes to resemble the more arbitrary distinction between "remunerated life and non-remunerated life" (2004, 103). The wages for housework perspective on the social factory demystifled both work and family by engaging some of the political-economic, ethical, and gendered discourses that undergird both spheres and promoted the cognitive mapping of the relations among work's various sectors. The demand for basic income has the potential to accomplish something comparable, although shifting the focus of its analyses from the Fordist to the post-Fordist social factory. Akhough its pedagogy is less clearly inscribed in the very language of the demand than the slogan "wages for housework' the demand for basic income nonetheless presumes an analysis of the political economy of the contemporary wage system, and to engage with the demand requires a reconsideration of its standard rationale. Rather than register the fact that some workersnamely, those performing unwaged domestic workare not now adequately included in the wage system, the demand for basic income points toward an even less reliable determination of who is and who is not included. The demand for basic income extends the insight of the wages for housework perspective that an individual's income depends on a network of social labor and cooperation broader than the individual wage relation (see Robeyns 2001, 84-85). Whereas the demand for wages for housework intended to expose the dependence of waged work on household-based relations of reproduction, the demand for basic income entails, as Ailsa McKay and Jo Vanevery observe, "an implicit recognition that all citizens contribute to society in a variety of ways:' including contributions "that may or may not have monetary value or even be measurable" (2000, 281). The demand for wages for housework sought to expose some of the inadequacies of the relationship between work and income by imagining what it might take to repair the wage system; the demand for basic income's proposal to break the link between work and income highlights the arbitrariness of which practices are waged and which are not.24 A major difference between the two demands is that whereas the demand for wages for housework served better as a critical perspective on the wage system than as a concrete proposal for reform, the demand for basic income offers both a critique and a constructive response. As a reform, basic income could help address several key problems of the post-Fordist US political economy that renders its wage system unable to function adequately as a mechanism of social distribution. These include the increasingly inadequate quantity and quality of waged labor manifest in high levels of unemployment, underemployment, and temporary and contingent employment, as well as the problem noted in chapter I of measuring individual contributions to increasingly collective and immaterial labor processes. The demand for basic income poses a critique but also provides a remedy: reducing our dependence on work. The demand for basic income presumes and evokes a critical perspective not only on the relationship between income and work, but also on the relationship between income and family membership. To recall our earlier discussion of wages for housework, as a perspective that demand tried to make visible the interdependence between the wage system and the institution of the family. The family is not a separate sphere, but part of society's economic apparatus. The family and its ideology help to obscure the costs of productive labor by privatizing, feminizing, and naturalizing much of the work involved in its reproduction. The problem is that neither the wage system nor the institution of the family is able to meet the needs of those individuals whose forms of productivity and intimacy do not line up with such restrictive institutions of social cooperation and economic distribution. One of the advantages of basic income is that, as McKay and Vanevery point out, it would be distributed to individuals irrespective of family membership or household form (2000, 281). In this way, the demand refuses to privilege either work or family as institutions on which an individual must depend if he or she is to secure the necessary means to support a life. Once again, the advantage of basic income is that it can both generate critical perspectives and offer an effective policy change. Whereas the wages for housework perspective sought to expose the link between the wage system and the family, as many have observed, its achievement risked preserving the relationship. As a perspective, the demand for basic income raises questions about whether narrow definitions of either work or family can or should suffice as principles governing the allocation of income (see McKay and Vanevery 2000, 268); as a concrete reform, it could ease the economic strain that can compel individuals to participate in both waged work and family membership. As Carole Pateman notes, "a basic income has the potential both to encourage critical reassessment of the mutually reinforcing structures of marriage, employment and citizenship, and to open the possibility that these institutions could be re-made in a new, more democratic form" (2006, iio). The demand for basic income thus recalls and amplifies both the antiproductivism and the antifamilialism of the wages for housework perspective. As a means to challenge at once the work ethic and the family-values discourse with which it is linked, this demand is reminiscent of an earlier demand for basic income that was advanced within a movement often cited in the wages for housework literature as a source of inspiration. The welfare rights movement, in both the United States and England, was another "revolt of the wageless" that the wages for housework authors found instructive (see, for example, Edmond and Fleming '975, 9; Cox and Federici 1976, 12). The demand for a basic income was in fact a key tenet of the US National Welfare Rights Organization in the 196os and 197os. Like advocates of wages for housework, the organization attempted to gain recognition for the labor of parenting while at the same time refusing the work ethic's praise for and privileging of work. Eileen Boris explains that the organization "recognized the necessity of not merely expanding the definition of work to embrace the unpaid labor of care giving or motherwork, but of refocusing the debate from work to income" (1999, 37). These activists were, Felicia Kornbluh (1997) argues, animated less by the notion of a right to work than by a right to consumption predicated upon an adequate level of income. As an effort to secure an income independent of wages, the demand for basic income registers the refusal of an ethics that enforces dependency either on marriage or the wage relation; indeed, the demand calls into question the adequacy of any ideal of social reciprocity that is reduced to a series of individual contracts. BASIC INCOME AS PROVOCATION As a perspective, a demand encourages critical reflection on the present order of things: what are the problems the demand seeks to address, and what is the rationale for the solution it puts forward? As a provocation, a demand points toward the future: what would be different if, for exam- ple, wages were paid for housework, or income were provided irrespective of work or family membership? As a mode of provocation, the collective practice of demanding should be understood also as a constitutive event, the performative force of which inevitably exceeds the scope of the specific reform. There are a number of different ways to approach basic income as a provocation to something new. I want to touch, very briefly, on two that bear interesting resemblances to the earlier discussion of the demand for wages for housework: basic income as a provocation to freedom and as a provocation of desire. As for the first of these, although the demand for basic income can certainly be seen as a means to reduce inequality, it can also be understood as an invocation of the possibility of freedom. By "freedom" I mean neither individual self-sovereignty nor libertarian license,25 but rather what the wages for housework tradition envisioned as a condition of collective autonomy: freedom as the time and space for invention. Basic income can be demanded as a way to gain some measure of distance and separation from the wage relation, and that distance might in turn create the possibility of a life no longer so thoroughly and relentlessly dependent upon work for its qualities. Therefore, we might demand a basic income not so that we can have, do, or be what we already want, do, or are, but because it might allow us to consider and experiment with different kinds of lives, with wanting, doing, and being otherwise. The demand for basic income could also be an occasion to contemplate the shape of a life beyond work, the kind of freedom that, as Marx speculates, "begins only when labour determined by necessity and external expediency ends' in a sphere of existence that lies "beyond the sphere of material production proper" (1981, 959). The demand can serve thus as a provocation to imagine the possibilities of a postwork alternative in which the structures, relations, values, experiences, and meaning of work might be substantially refigured. But perhaps the most provocative aspect of the demand for basic income today is its anti-asceticism. Indeed, it is worth noting that in debates about basic income, cost is not necessarily the primary point of contention.26 Rather, it is the ethics of the demand that, often seems to generate the most discomfortspecifically, over the way the demand is seen to denigrate the work ethic and challenge ideals of social reciprocity that have been so firmly attached to the ideal of the labor contract.27 Here too the demand for basic income echoes the demand for wages for housework: both speak to the possibilities of subjects rich in desires and needs. As a provocation of desirefor more money, more time, more freedomthe demand for basic income, like the demand for wages for housework, sets itself apart from so many other approaches to political claims making. Rather than preach the ethics of thrift and savings, the politics of concession, or the economics of sacrifice, the demand for basic income invites the expansion of our needs and desires. In contrast to the more familiar styles of political analysis and strategy that revere work and decry consumerism, it rejects the usual prescription that we should work harder and want less. On the contrary, the demand is excessive, defying what are proclaimed to be reasonable limits on what we should want and demand. By challenging the link between individual production and consumption, by refusing the notion that waged work is the only legitimate means of access to even a minimal standard of living, the demand for basic income points in the direction of a life no longer subordinate to work. On the one hand, this refusal of asceticism may render the demand more difficult to achieve and, in that sense, limit certain aspects of its power as a perspective and provocation. On the other hand, to anticipate an argument I will pursue in chapter 5, the demand is also compelling because it departs from those strictly productivist values that link the worth of individuals to their commitment to work and that tether access to income to its performance. Precisely where the demand fails to pass muster with a model of political calculation sutured to the present may be where it can succeed in sparking the political imagination of, and desire for, a different future. Using our earlier reading of the wages for housework literature as a model for our consideration of the demand for basic income allows us to recognize the latter demand as not merely a policy proposal but a perspective and a provocation, a pedagogical practice that entails a critical analysis of the present and an imagination of a different future. What makes this demand a worthy successor to the 1970S demand for wages for housework has to do with its advantages as a perspective and provocation, but also as a reform. Indeed, it is arguably a better vehicle by which to advance some of the key goals of the earlier movement: as a perspective, it can challenge both productivist ethics and family values and provoke the possibility of a social form that no longer privileges these now-dominant regimes of economic production, social cooperation, and political order. The potential of the demand to be both epistemologically and ontologically generative ensures the value of advancing it despite the fact that its success in the short term is a long shot. What increases its worth as a successor project is that as a practical reform, basic income offers tangible benefits to a broader constituency than the housewives who were the focus of the earlier demand. In terms of the two critiques of the demand for wages for housework discussed abovenamely, that the gender division of labor would be further entrenched by the payment of this wage to housewives, and that the integrity of the wage system would be upheld rather than contested by rewarding more forms of work with wagesthe demand for basic income is a more viable alternative. By proposing to award the income universally to individuals and thus lessening the dependence of income on work, basic income not only recognizes but offers a response to the inability of both the wage system and the institution of the family to serve as reliable mechanisms of income distribution.

#### Post-work causes the libidinization of life – compulsory labor is psychological torture and a denial of your true self and meaning outside of capitalistic production

Seyferth 19 [Peter Seyferth, university lecturer on political theory and philosophy with a MA and a PhD, 2019, "Anti-Work", Taylor & Francis, https://www.taylorfrancis.com/chapters/edit/10.4324/9781315619880-31/anti-work-peter-seyferth]/Kankee

Neo-Marxist Anti-work André Gorz stands at the vanguard of contemporary neo-Marxist anti-work thought. He analysed the transformation of work regimes in developed countries and argued that the neo-liberal order has already abolished work on a massive scale, but in a way that is exploitative and dominating, forcing everyone to fight each other for vanishing employment. In contrast to the productivist tradition of orthodox Marxism, he claims that ‘[i]t is not this abolition we should object to, but its claiming to perpetuate that same work, the norms, dignity and availability of which it is abolishing, as an obligation, as a norm, and as the irreplaceable foundation of the rights and dignity of all’ (Gorz, 1999: 1). Taylorism/Fordism had to totally repress the self-organisation, ingenuity and creativity of the workers because it understood them as sources of rebellion; the new regime of Toyotism/Post-Fordism treats these as ‘a resource to be developed and exploited’ (30). Hence, everyone is forced to sell their personality, to commodify their selves. One has to show a particular state of mind, a willing disposition: ‘eagerness’ (43). The professionalisation of pleasant and attentive attitudes ultimately poisons the arts of living because now even a nice smile falls under the suspicion of hypocrisy: it could be the ‘commercial smile’ or ‘standardized, superficial bonhomie’ of salespersons (70–71). Gorz’s criticism points not just to a hollow lifestyle, but to a profoundly political struggle for power because ‘the unfettered power capital has assumed over labour, society and everyone’s lives depends precisely on “work” – not the work you do, but the work you are made to do – retaining its centrality in everyone’s lives and minds’ (53, see also 58). So, capitalism has an ‘ideological hold on people’s minds’ by refusing income from activities that are not ordered and paid for by others, and thereby persuading people of the ‘imperative need to work’ (72). Other neo-Marxist thinkers like Nick Srnicek and Alex Williams or Michael Hirsch build on and expand Gorz’s critique. Typically, they emphasise the odd effects of high productivity. This should make classic demands of the left achievable (provided that distribution will be changed by reform or revolution), but it, in fact, worsens the workers’ plight by increasing the threat of unemployment. The liberal notion of ‘freedom of contract’ has become an absurdity as neo-liberal state policies have transformed welfare to workfare – basic social rights have been supplanted by compulsory work (Srnicek and Williams, 2015: 1–2, 93, 100–104; Hirsch, 2016: 10, 14, 40, 76). Hirsch exposes contemporary work society to be a Platonic order. In his Republic, Plato (1991: 111 [433a]) claims that in the just state, ‘each one must practice one of the functions in the city, that one for which his nature made him naturally most fit. […] justice is the minding of one’s own business and not being a busybody’. This is precisely the opposite of Gorz’s (1999: 73) alternative to the wage-based society: ‘multi-active life’. But according to Hirsch, today’s developed countries follow the Platonic idea: everyone has to have a work position they identify with. The economic existence and social status of the individual depends on a lifelong full-time job. Since there are not enough such jobs, the Platonic order makes individuals accept economically meaningless occupations, sometimes by threat of a debasing exclusion, and sometimes by sheer force. On the one hand, this is necessary to reproduce the hierarchy of classes, professions and genders. On the other hand, this work does not add value to the economy (Hirsch, 2016: 16–17, 20, 46, 59–66, 74). So, it is purely ideological: a cultural hegemony upheld by an alliance of capital, unions, social democracy, state apparatus, media and academy. Progressives should re-enter the fight over the definition of reality, over what we take for granted. Hirsch theorises the hegemony of work with Pierre Bourdieu’s notion of ‘symbolic violence’ (22, 27–29, 68–69, 104, 177); Srnicek and Williams stick to Antonio Gramsci’s older but similar concept of ‘hegemony’ for their analysis. ‘The fact that so many people find it impossible to imagine a meaningful life outside of work demonstrates the extent to which the work ethic has infected our minds’ (Srnicek and Williams, 2015: 124). This has an almost religious quality: ‘The central ideological support for the work ethic is that remuneration be tied to suffering’ (125). They concentrate on the genesis of neo-liberalism’s hegemony, from the Mount Pelerin Society through networks and ideological infrastructures to palpable policies and politicians – and propound a similar counter-hegemonic strategy to eventually arrive at a world without work (52–67). The neo-Marxists discussed here approve of imagining future alternatives to work society, but do not give many details. Gorz (1999: 77, 78) aims at ‘the multi-activity-based society’ where ‘everybody will engage in a range of different activities and modes of membership of the society’. Srnicek and Williams (2015: 85–86) want a ‘post-work world’ that is ‘not a world of idleness; rather, it is a world in which people are no longer bound to their jobs, but free to create their own lives’. Hirsch’s (2016: 251, 43) ‘utopia of an egalitarian and variegated society’ promises ‘an equal freedom of, in, and to work. Ultimately, it is about free access to and equal recognition of all forms of work’.3 However, the neo-Marxists rather focus on transformative strategies, especially the role of the state in the counter-hegemonic emancipation from work. Since this is a fight about the definition of reality and since the state (as a ‘bank of symbolic capital’) produces and enforces categories of thought, it is crucial how the state understands its function, argues Hirsch (27–32). In the case of the creation of jobs being perceived as the state’s mission, the state is made undemocratic; in the other case, if the mission is to eliminate useless work, a lot of reforms are obviously necessary: wealth, working hours and skills have to be redistributed justly; social security, individual freedom and democratic sovereignty have to be provided without force. From this follows voluntary social self-organisation (9, 11, 103, 105). Similarly, Gorz (1999: 100, 21) has already advocated the state to create free space and thus the possibility of independent activity, financed by some sort of Tobin tax that presupposes the state’s willingness to overpower financial capital. And Srnicek and Williams (2015: 108) demand ‘non-reformist reforms’ that push capitalism beyond its acceptable parameters; such reforms would be, for example, more state investment, research devoted to technologies that replace workers, higher minimum wages, support for labour movements, reduction of working hours, etc. (112–14). These neo-Marxists agree on one central state policy to be the key for emancipation: a UBI. It is a means ‘to free them [the recipients] from the constraints of the labour market. The basic social income must enable them to refuse work and reject “inhuman” working conditions’ (Gorz, 1999: 83). Its significance ‘lies in the way it overturns the asymmetry of power that currently exists between labour and capital’ because with it, workers ‘have the option to choose whether to take a job or not […] A UBI therefore unbinds the coercive aspects of wage labour, partially decommodifies labour, and thus transforms the political relationship between labour and capital’ (Srnicek and Williams, 2015: 120). It saves the workers from blackmail by the state and capital, grants more autonomy regarding time and lifestyle, enables activities that make sense but are not profitable, changes power relations between employers and employees, and ultimately quits the Platonic order (Hirsch, 2016: 101, 111–17). Alas, the UBI is a tool that could serve the capitalists as well (and has been demanded by neo-liberals), as all four neo-Marxists hasten to concede. If it is set below the breadline, its recipients are still forced to accept all kinds of jobs – then it is just a subsidy for capitalists who offer jobs no one would do without compulsion (Gorz, 1999: 82; Srnicek and William, 2015: 119; Hirsch, 2016: 102). It remains unclear how this danger could be averted; one may only guess that the neo-Marxists assume that the government that fulfils their demands will withstand all well-funded lobbying activities and elections from then onwards. This is where the neo-Marxists most obviously deviate from anarchists. Although they appreciate the absenteeism and even rebelliousness of workers who hate work and ennoble the surplus population with titles like ‘revolutionary subject’ or ‘vanguard’, it is ultimately the state, preferably represented by left-wing populist politicians, that has to set things right (Gorz, 1999: 40, 63–64, 79, 110; Srnicek and Williams, 2015: 47, 114, 156–60; Hirsch, 2016: 230–1, 243–4). Srnicek and Williams (2015: 11, 22–23, 26–29, 32, 35) explicitly position themselves against the ‘contemporary anarchist-tinged politics’ of the contemporary left. This tendency of ‘horizontalism’ or ‘folk politics’ as exemplified by Michael Bakunin and Peter Kropotkin in the past and by small anarchist movements analysed by Uri Gordon today face a barrage of Srnicek’s and Williams’s criticism: it is too exclusive, too small, too harmless – in their view ‘evidence shows that hierarchical organisations are crucial in defending movements against the state’ (33). But, from an anarchist perspective, the fault of classical Marxism is that these hierarchical organisations tend to become wielders of state power themselves; this is why, so many neo-Marxists are quite critical of taking power nowadays. Following Gorz, who, in passing, mentions some Proudhonists and Paul Goodman but ultimately claims the indispensability of the state (1999: 84, 110), Hirsch (2016: 227) argues for a ‘radically anarchic position of individual autonomy’ without actually embracing anarchism. Almost like Kropotkin, he demands to politically abolish the divisions of work (both in the paid and unpaid spheres), the cause of most forms of social and gender hierarchies (240). For him, criticism of the work society is only emancipatory if it is also a radical criticism of domination (247). All this is typical for anti-work anarchists, but for Hirsch, the state must not be destroyed, but appropriated in a radical democratic way (251). Some neo-Marxist anti-work arguments are wholeheartedly hostile to the state, notably the Krisis group’s Manifesto against Labour (1999). Here, the state is seen as a derivative of labour; a state-controlled UBI is thus rejected; generally, ‘the opponents of labour don’t want to take the control centres of power, but want to switch them off’ (Krisis, 1999). The Krisis group developed its theory on the grounds that ‘Marxism itself was not sufficiently radical in its critique’ (Trenkle, 2001). Krisis dismisses the classical class struggle of unions and parties whom it perceives to be part of the ‘labour camp’ (i.e. those that fight for work). On the one hand, Krisis (1999) promises that ‘[m]arket and state, institutions (once) alienated from human society, will be replaced by a graded system of councils, from town district level to the global level, where associations of free individuals will decide about the flow of resources in letting prevail sensual, social, and ecological reason’. For Wildcat, a syndicalist circular with tendencies towards communism, this is a sign of Krisis’s individualist anarchism (H., 1999). On the other hand, for Krisis (1999), ‘[o]nly an explicitly formulated critique of labour along with a corresponding theoretical debate could bring about a new public awareness; the latter being the indispensable prerequisite for the constitution of a social movement that puts labour critique into practice’. This is a sign of Krisis’s Leninism, and its supposed intellectual superiority over the masses who lack the proper consciousness, according to Wildcat’s anonymous author (H., 1999). Anarchist Leninism seems to be a contradiction in terms. Let’s have a look at those enemies of work that actually self-identify as anarchists to shed some light on the differences between anti-productivist neo-Marxists and anti-productivist anarchists. Anarchist Anti-work The two most obvious differences between the neo-Marxists and the anarchists are the latter’s rejection of hierarchical organisations (like the state) and their focus not on theory but on resistant action. Surely, anarchists develop theories, too. But by consciously devaluating even their own theories, they shun the dogmatic scholasticism they see in Marxism. From this follows a multiplicity of anarchisms that are to some extent internally contradictory. Anti-productivism is notably strong in the primitivist and insurrectionist currents of anarchism, which are also notoriously hostile to academic theory. John Zerzan (1999: 133) stresses that ‘our many compromises and accommodations with a grisly world are the real field of our effort to break free, more so than merely stating our ideas’. In this field, Marx has failed totally, both personally (with his ‘ruling-class mentality’ bringing misery to his family; 138, 133) and politically (with the ‘gradualist, collaborationist, and highly statist’ nature of his demands, essentially accepting the misery of capitalist development; 135). Zerzan’s main point of criticism of work, capitalism and civilisation itself is the alienation inherent in them. Alienation is also a central term of Marxist theory. But ‘Marx’s overriding concern with externalities – principally economic crises, of course – was a trademark of his practical as well as theoretical approach; it obviously reflects his slight regard for the subjectivity of the majority of people, for their potential autonomy, imagination, and strength’ (138). So, Zerzan focuses his own observations of work precisely on these bottom-up resistance practices that are typically devoid of theory. In a series of articles published from 1974 to 1986, he tells the story of unruly struggles against industrialism. From the mid-eighteenth century onwards, artisanal freedom is pressurised by the rising capitalist factory system. Against this, ‘[a]bsenteeism, as well as turnover, […] was part of the syndrome of striving to maintain a maximum of personal liberty’ (100). The height of this resistance was ‘the widespread revolutionary movement’ of the machine breaking Luddites (105) that was almost anarchist: ‘the Luddites organized themselves locally and even federally, including workers from all trades, with an amazing, spontaneous coordination. schewing an alienating structure, their organization was neither formal nor permanent. Their revolt tradition was without a center’, although it ‘was not a completely egalitarian movement’ (109). Zerzan draws a sharp contrast between, on the one hand, machine breakers, rioters and insurrectionaries ‘who mounted the most extreme forms of opposition’ (117) and, on the other hand, the labour movement organised in parties and unions. The latter helped shape the factory model of society (that itself reproduced the prison model), fulfilling the factory owners’ wish for ‘a more subdued populace’ (118, 129). Hereby, ‘the union became the effective agency for suppressing workers’ direct action against speed-up or other grievances’ (181). By reducing class struggle to wage bargaining, the unions separate the workers from autonomous control over the working conditions. Because they are mainly interested in member fees and often strive to control the workers by ‘closed shop’ arrangements (so that, just like the bosses, union officials can fire unruly workers), unions are actually instruments of domination and not the right place to start fighting. Zerzan is sure most people feel deep inside that ‘the falseness of trading away one’s life in order to purchase things is a transparently barren death-trip’ (233). He cites a great many examples of disenchantment, withdrawal and resistance to work: absenteeism, turnover, employee theft, illness, drinking, drug-taking, suicide, stress, job burnout, misuse of on-the-job time, sabotage (208, 213, 217, 219, 222, 223, 228, 231–2, 245, 246, 248, 256, 258). But Zerzan’s criticism goes much farther. Disenchantment, withdrawal and resistance also turn against everyday life: declining civic virtue, increasing shoplifting, an immense prison population, tax cheating, child arson, insomnia, emotional disability, loneliness, depression, anorexia, bulimia, gambling, refusal of literacy – an intentional aversion to the whole of modern life (217, 246, 247, 257–8). ‘The possibility that the impoverishment of daily life might even render work relatively satisfying, due to the vacuum of substance elsewhere, is rendered unlikely by technology’s progressive degradation of work. There is no area of authenticity, no place to hide, and no one can miss this commonplace’ (257). Zerzan was himself a union official in the 1960s. His turn away from organised leftism and towards a broader anti-work and even anticivilisation attitudes was inspired by anarcho-primitivist Fredy Perlman’s (1983: 37) conviction that ‘[l]abor is always forced labor’ and that the Stone Age was some kind of Garden of den; it was civilisation that turned ‘the Garden into a forced-labor camp’ (258). Zerzan (1994: 16) adopted this view that ‘life before domestication/agriculture was in fact largely one of leisure, intimacy with nature, sensual wisdom, sexual equality, and health. This was our human nature, for a couple of million years, prior to enslavement by priests, kings, and bosses’. The enslavement has taken the form of division of labour, which is deeply alienating. ‘The relative wholeness of pre-civilized life was first and foremost an absence of the narrowing, confining separation of people into differentiated roles and functions’. This social stratification has been a central idea of ‘key ideologues of civilization’ since Plato (147). But the division goes even deeper – its initiation coincides with the ‘dissolution of the original unity between humanity and nature’ (Zerzan, 1999: 37). This split makes humans into alienated beings that have to use abstract symbols (time, language, number, art, etc.) instead of directly interacting with their environment. It was the implementation of agriculture that brought this estrangement to humanity. Agriculture is ‘the end of life as mainly sensuous activity, the embodiment and generator of separated life. Artificiality and work have steadily increased since its inception and are known as culture: in domesticating animals and plants man necessarily domesticated himself’ (73). Before the neolithic revolution, work was inexistent. Since then, life quality deteriorated as repression and alienation increased, especially in the course of industrialisation. Because the ‘[d]ivision of labor embodies, as an implicit purpose, the control and domination of the work process and those tied to it’ (118), it has to be overcome. For Zerzan, this does not only entail the abolition of work, but of all alienation and thus of civilisation itself. It remains somewhat unclear how this may actually be done. Bob Black (1991), in his 1986 essay The Abolition of Work, proposes a ‘ludic revolution’ that creates ‘a new way of life based on play’. In the 1980s, Black (1992: 11) learned of ‘the idea of zerowork. I’d already been brought close to it by such writers as Fourier, Morris and Kropotkin when I discovered John Zerzan’s studies of work refusal’. Like Perlman and Zerzan, Black (1991) is convinced that ‘work is forced labor, that is, compulsory production’, which is ‘the source of nearly all the misery in the world’. Work degrades workers by discipline (‘surveillance, rotework, imposed work tempos, production quotas, punching-in and out, etc.’), and makes them dumb and habituated to hierarchy, even in their free time (which is ‘mostly devoted to getting ready for work, going to work, returning from work, and recovering from work’). Work, in the end, is deadly – it is ‘mass murder or genocide. Directly or indirectly, work will kill most of the people who read these words’. Black attributes most deaths by cars, industrial pollution, alcohol, drugs, cancer and heart disease to work, because ‘[e]ven if you aren’t killed or crippled while actually working, you very well might be while going to work, coming from work, looking for work, or trying to forget about work’.4 Work steals workers most of their time, even if nothing valuable is produced in working time. Actually, ‘most work serves the unproductive purposes of commerce or social control’. Therefore, useless and pernicious work has to be cut down massively. Insofar as work serves useful purposes, it has to be transformed into ‘a pleasing variety of game-like and craft-like pastimes, indistinguishable from other pleasurable pastimes except that they happen to yield useful end-products’. Black takes inspiration from Charles Fourier. Most necessary and useful activities are attractive to many people – at least from time to time, as long as they are done without compulsion but in enjoyable circumstances. This also evokes William Morris’s utopia of work turned into artisanal handicraft (in his News from Nowhere).5 Black agrees with Zerzan that technology does not save labour: ‘When productive technology went from hunter-gathering to agriculture and on to industry, work increased while skills and self-determination diminished’. But Black does not want to do away with all technology. Contrary to the neo-Marxists cited earlier, he agrees with Morris that technology should play just a small role in this transformation. Yet, Black’s quite optimistic outlook has a different flavour from Morris’s: ‘An optimal sexual encounter is the paradigm of productive play. The participants potentiate each other’s pleasures, nobody keeps score, and everybody wins. The more you give, the more you get. In the ludic life, the best of sex will diffuse into the better part of daily life. Generalized play leads to the libidinization of life. Sex, in turn, can become less urgent and desperate, more playful’. A series of anti-work articles written by the insurrectionalist anarchist Alfredo M. Bonanno (1987–1995) are much bleaker and more militant. He distances himself from most other anti-work activists and theorists, asserting he does ‘not feel any nostalgia for lost professionalism’, being ‘even less interested in elaborating libertarian alternatives to grim factory work or intellectual labour’, and going so far as to decry ‘the abolition of work or its reduction to the minimum required for a meaningful happy life’ (Bonanno, 2009: 6). Not even the substitution of work with play satisfies Bonanno, because ‘none of that escapes the essential rules of work seen in terms of the global organization of control’ (8). Traditional anti-work strategies have mostly been recuperated by capitalism, so they ‘have now become normal procedures for capital itself’ (9). Like Zerzan, Bonanno decries alienation (14) while deeming Marxist ideas useless (27) and criticising the unions for their participation in business decisions, whereby they abandon more militant activism and essentially pursue ‘complete social pacification’ (45). ven self-managed co-ops are too peaceful and impotent and therefore not revolutionary enough (17, 38). ‘So we see that work cannot be abolished progressively: we need to approach the problem in a destructive manner’ (8). Bonanno mainly discusses two strategies: sabotage and bank robbery. They are destructive, but not even they suffice: sabotage cannot destroy the worker identity, and robbery can become a full-time job – in Bonanno’s experience (as a convicted robber talking to other prison inmates), the crucial problem is to know what to do with all the money (10–11). So, ‘the essential part of any project to destroy work is creativity taken to the maximum possible degree’ (10). Bonanno calls this ‘projectuality’ (‘progettualità’) – having a ‘project in terms of life’, or rather inventing this project ‘by reflecting upon what one wants to do with one’s life and finding the necessary means to realise it, without working’ (11). Wolfi Landstreicher (2003) explains ‘projectuality’ as a prefigurative autonomous direct action that is explicitly aggressive, an ‘active refusal of alienated existence’ and ‘the reappropriation of life’. Referring to Bonanno, Landstreicher gives a definition: ‘anarchist projectuality is the practical recognition in one’s life that anarchy is not just an aim for the distant future, an ideal that we hope to experience in a far away utopia. Much more essentially, it is a way of confronting life and struggle, a way that puts us at odds with the world as it is. […] When we make the choice to cease to be a cog, when we make the choice to break the machine rather than continuing to adjust it, passivity ceases and projectuality begins’. Arguably, the publications of the CrimethInc. ex-worker’s collective combine such a projectuality with Black’s ludic revolution. CrimethInc. positions itself at the post-left end of the anarchists’ spectrum. The root of the collective is anarcho-punk (notably the band Catharsis). According to Thompson (2004: 109), CrimethInc. ‘began as Brian D[ingledine]’s personal zine, Inside Front, and began operating as a collective only in 1996’. In Inside Front, personal accounts of travelling and dropping out are prevalent. Dingledine (1996) explains why he refuses to work: ‘When I don’t work for them, they don’t get to use my labor to perpetuate the status quo. When I don’t receive an income from them, I don’t have capital to give back to them for them to use to perpetuate the status quo. And most of all, my time and energy are mine to be used to fight against them, rather than to support present conditions’. Similarly, the later CrimethInc. collective asserts that because it is the combined labour and consumption of all workers and employees that powers the system, resistance cannot be a part-time hobby of people who work full-time to maintain it (CrimethInc., 2004: 8). The revolutionary purpose is not just quitting jobs, but ‘deserting and ultimately destroying the class system itself. […] a universal rejection of all possible positions within the social order, in order to create classless communities’. So, CrimethInc.’s class struggle is a struggle against class as such (9). Consequently, ‘[w]e shouldn’t base our solidarity on shared attributes or social positions, but on a shared refusal of our roles in the economy’ (CrimethInc., 2011: 250). Instead of dropping out, it would sometimes be better to use the resources of your position for revolutionary purposes and interrupt the system in your position until they fire you (349). ‘Not working is only half the battle, and not the important half, at that. The real question is what you do instead. […] You only discover what interests you by engaging with the world, and in capitalist society, employment and consumption can seem like the only ways to engage. As you cut down on these, replace them immediately with new projects, with the things you dreamed of doing when you didn’t have a moment free’ (CrimethInc., 2005: 582). This could be volunteering in community groups (or inventing them), making a project out of enjoying oneself, learning, exploring, following ideas – and being happily and proudly unemployed, so as to radicalise the community one comes from (582–3; cf. CrimethInc., 2004: 9). The CrimethInc. collective criticises (neo)-Marxism for making radicals try to look like the mainstream, so as to influence it in revolutionary ways; dropping out, by contrast, creates many alternative ways of life, thereby challenging the dominant order (CrimethInc., 2006: 13). To have energy to fight the system, one needs food, shelter and other resources, of course. CrimethInc.’s most popular strategy for non-symbolic assaults on the system and no-work sustenance at the same time is theft from corporations. Shoplifting (as well as dumpster diving and squatting) not only rescues resources, it also gives the feeling of freedom and power; it is an attack on consumerism, and it is better than boycotting, because it directly harms the corporations (CrimethInc., 2001: 237–9; cf. Dingledine, 1996). But stealing individually without revolting will not interrupt the status quo – so stealing should be part of building community (CrimethInc., 2011: 276, 278). The hope is that fighting and stealing and living determinedly and desirously might be infectious, so others join in to reclaim the resources of the society, ultimately leading to transformation (CrimethInc., 2001: 256). The collective invites the reader: ‘Join us in making the “revolution” a game; a game played for the highest stakes of all, but a joyous, carefree game nonetheless!’ (192) This sounds quite hedonistic, but at least in later publications, CrimethInc. does not demand that everything is fun or easy; fighting work is hard work (CrimethInc., 2011: 38). And in spite of the individualistic – often even Stirnerite – flavour of much CrimethInc. literature, the struggle against work is one that requires massive support networks and connections between disparate social circles. These should not be organisations with homogenous membership, but coalitions of solidarity between anticapitalist people of all walks of life (CrimethInc., 2004: 9).

#### Work is capitalistic domination and a denial of dignity

Seyferth 19 [Peter Seyferth, university lecturer on political theory and philosophy with a MA and a PhD, 2019, "Anti-Work", Taylor & Francis, https://www.taylorfrancis.com/chapters/edit/10.4324/9781315619880-31/anti-work-peter-seyferth]/Kankee

Rejection of the work regime is neither ubiquitous amongst nor confined to radicals. But those that radically turn against work argue that it is a relationship of exploitation, submission and alienation that has to be abolished. Indeed, there is a long history of worker resistance against drudgery, starting from Luddite sabotage, later tamed by unions, nowadays appearing more silent as absenteeism. There are two complementary radical frameworks of anti-work. On the one hand, neo-Marxists criticise work ideology and demand a universal basic income (UBI) from the state. On the other hand, anarchists reject the state and focus on direct resistance action against work. One of the core elements of modern domination is work. Without work, capitalism and, arguably, the state would be impossible. Anti-work radicals argue and struggle for the abolition of work, which is, in their view, a precondition for freedom from domination. Work is a praxis and a power machine that acts on intersecting social identities, as Fineman (2012: 3) has observed: ‘Today, pecking orders of who does what kind of work remain in all parts of globe, heavily influenced by social class, education, wealth, gender, race, age, or ethnicity’. Alongside the paid jobs that appear in the Gross National Product and are often held by privileged people (mostly skilled, white, healthy males), there are several other kinds of work that are even more frustrating, dangerous and demeaning while yielding less resources (or none at all) and thus have to be done by underprivileged people: domestic labour, voluntary work, black economy (13). Although each of these activities could theoretically be done due to its intrinsical value – if the respective worker happens to attribute such a value to the activity – most work is regarded as means for some external good: as a necessary evil. Ethical attitudes towards work vary wildly from appreciation to rejection (5), so to motivate people to work, work has to meet the needs of people (10–12). Since many jobs fail to do so, there have to be mechanisms to force people to work. The easiest way is to deprive those that do not work from what they need and want, be it food and shelter or status, power and dignity – although even those that do work often lack many of these essential goods. If one must work to survive, one can and will easily be exploited. Emancipatory radicals deal with the (at least potential) evils of work in quite different ways: some aim at improvements within the sphere of work; others strive for the abolition of work itself. The better-work/ anti-work cleavage cuts across the usual dividing lines of radical traditions. In the following, I will first show the ambiguity of nineteenth- and early twentieth-century radical thinkers towards work. Second, I will sketch the transformation of work regimes in the second half of the twentieth century and assess worker’s everyday resistance to it. Third, I will present the theories of several anti-work neo-Marxists. Finally, I will highlight central points of anarchist anti-work agitation. My aim is to shed light on the differences of anti-work radicalisms while making plausible why one might want to abolish work. Radical but Ambiguous Stances Over Work in the nineteenth and early Twentieth Centuries The most widely known critic of capitalism, Karl Marx, stated in his early manuscripts of 1844 that the oppression of workers is not just a case of economic exploitation, but also of human estrangement. On the one hand, the worker is alienated from the product of his work (it does not belong to him but to the capitalist). But on the other hand, ‘estrangement is manifested not only in the result but in the act of production, within the producing activity, itself’. Thus, labor is external to the worker, i.e., it does not belong to his intrinsic nature; that in his work, therefore, he does not affirm himself but denies himself, does not feel content but unhappy, does not develop freely his physical and mental energy but mortifies his body and ruins his mind. (Marx, 2009: 30) Work seems to be a vile thing. One might expect Marx, being an infamous radical, to reject work altogether. But Marx’s theoretical attitude is quite ambiguous. In the third volume of his Capital (posthumously published in 1894), he indeed seems to contrast freedom with work: ‘the realm of freedom actually begins only where labour which is determined by necessity and mundane considerations ceases’. Too bad that this is impossible, for Marx. People will have to work ‘in all social formations and under all possible modes of production’, including communism; there always remains ‘a realm of necessity. Beyond it begins that development of human energy which is an end in itself, the true realm of freedom, which, however, can blossom forth only with this realm of necessity as its basis’. The best thing to hope for is thus a ‘shortening of the working-day’, made possible by increased productivity (Marx, 2010: 571). Consequently, Marx’s most successful followers (Leninists and Social Democrats) sought to appear not only worker-friendly, but also work-friendly. They adhered to a strict work ethic and imposed it on their subjects whenever they took state power. The culmination of this was the Stakhanovite movement that tried to maximise production without cutting working hours from 1935 onwards, but already in 1879 August Bebel (1904: 275) was quite clear on this point: ‘So soon as society is in possession of all the means of production, the duty to work, on the part of all able to work, without distinction of sex, becomes the organic law of socialized society’. It is plausible to attribute Marxism’s fondness of forced labour not only to its inherently authoritarian utopian vision (where coerced work is one tool of the dictatorship of the party), but also to its official rejection of bourgeois lifestyle and social stratification. Indeed, the ruling classes of all times showed a distinct odium for work. This attitude may have been stronger with the nobility in antiquity and the Middle Ages than with the robber barons and misers of capitalism, but the unequal distribution of productive and leisured activity is similar. An example of the aristocratic devaluation of work can be found in Friedrich Nietzsche’s thought, especially from 1878 to 1882. For him, dirty work is a true vice that interferes with the desire for autonomy because it accustoms workers to repetitive petty goals that are easy to achieve. Insofar as it does so, work polices and subdues. Individuals become overwrought and their spiritlessness impedes cogitation. Thusly, idle time is experienced as boredom by workers; idleness is suitable only for artists and thinkers who need it for their creative work. Consequently, Nietzsche demands a robust division of humanity into two species: many ‘slaves’ with menial jobs, and a few freemen with aesthetic duties that do not occupy more than one-third of a day.1 This kind of attitude was not uncommon amongst anti-egalitarians. No wonder, then, that Bebel (1904: 275) cried: ‘The silly claim that the Socialist does not wish to work, that he seeks to abolish work, is a matchless absurdity, which fits our adversaries alone. Non-workers, idlers, exist in capitalist society only’. Similar ideas appear in Paul Lafargue’s famous 1880 essay The Right to be Lazy. The capitalist class appears to be ‘settled down into absolute laziness and demoralized by enforced enjoyment’. Unsurprisingly, ‘[t]he sight of the miserable conditions of life resignedly accepted by the working class and the sight of the organic degradation engendered by the depraved passion for work increased its aversion for all compulsory labor and all restrictions of its pleasures’ (Lafargue, 2010: 14). Like Nietzsche, Lafargue relies on ancient Greek and Roman thinkers (5, 21–22). But unlike Nietzsche, Lafargue was a socialist, and unlike Bebel, he harshly criticised the notion that ‘the proletariat, betraying its instincts, despising its historic mission, has let itself be perverted by the dogma of work’ (5). He contrasted the ‘noble savage’ that does not yet know work with Europe’s workers, ‘our miserable slaves of machines’, and identified work as ‘the cause of all intellectual degeneracy, of all organic deformity’ (4). In the end, Lafargue demanded that the proletariat ‘must return to its natural instincts, it must proclaim the Rights of Laziness’ (10), which in practice, means: ‘It must accustom itself to working but three hours a day, reserving the rest of the day and night for leisure and feasting’ (11). This would be made possible by technological progress. In a similar vein, Bertrand Russell, in his 1932 article In Praise of Idleness, speculated on the emancipatory possibilities of Taylorism (Frederick Taylor’s highly compartmentalised and deskilled division of labour which imposed strict factory discipline and was vividly exemplified in the Tramp’s experience on the assembly line in Charlie Chaplin’s 1936 classic, Modern Times): ‘If at the end of the war the scientific organization which had been created in order to liberate men for fighting and munition work had been preserved, and the hours of work had been cut down to four, all would have been well’ (1932: 554). All this is not far from Marx’s claim that with automation, the realm of necessity could be reduced – without completely abolishing work. What about those socialists that valued freedom over parties and discipline? The anarchists, too, have had ambiguous attitudes towards work. Like most other socialists of his time, Peter Kropotkin embraced technological progress and shorter working hours, e.g. in many articles published from 1886 to 1890. But to this question of efficiency, he adds the question of everyday joyous experience. Factories are needed (and possible) that are ‘so well managed that it would be a real pleasure to work in them, if the work, be it well understood, were not to last more than four or five hours a day, and if every one had the possibility of varying it according to his tastes’ (Kropotkin, 1907: 144–5). Work has to be agreeable, and this includes housework, because ‘a revolution, intoxicated with the beautiful words Liberty, quality, Solidarity would not be a revolution if it maintained slavery at home. Half humanity subjected to the slavery of the hearth would still have to rebel against the other half’ (154–5). Therefore, Kropotkin attacks the division of labour by gender and/or skill differences and calls for the abandonment of ‘any system of wages’ (212). To be entitled to the products one needs, it should be irrelevant how long or how qualified one has worked before; otherwise, a dividing line between new classes would be erected (205, 207, 210–13). The principle of distribution should not be wealth or work, but need; the work necessary for the satisfaction of needs should be minimised and as varied as possible (223). Using Adam Smith’s example of the smith that is ‘sentenced for life to the making of heads of nails’, Kropotkin illustrated the frustration, emptiness and disempowerment such divided work imposed on workers (232–3). His alternative vision entailed all people freely working half a day, spending the other half in artistic or scientific activities and retiring at the age of forty (Kropotkin, 1998: 187). This stands in stark opposition to Marx’s or Russell’s ideas.2 It must not be misunderstood as appreciation of idleness, though: Kropotkin (1907: 197–199) was sure that people hated and tried to avoid work only when it stifles them; to suppress this cause of laziness through the revolutionary restructuring of work in a meaningful and rewarding way would end idleness without the need to punish sluggards. Sadly, this emancipatory strategy towards necessary work and its avoidance has not always been followed by anarchist activists. Take as an example the CNT/FAI industrial policy during the Spanish Civil War 1936–39. Spain seemed to be an underdeveloped country to many Spanish anarcho-syndicalists who, consequently, glorified work as emancipatory and actively promoted industrialisation, as Michael Seidman (1991: 42) has observed. One of the CNT leaders, wartime Minister of the conomy Diego Abad de Santillán, ‘noted approvingly that Taylorization had eliminated the “unproductive movements of the individual” and had increased “his productivity”’ (45); he ‘underlined the necessity of eliminating “parasitism” and of providing work for all. Work would be both a right and a duty in revolutionary society, and he approved the old saying, Those who do not work, do not eat’ (46). This is quite the opposite of Kropotkin’s anti-Taylorist strategy; but in the civil war situation, a productive military industry seemed to be essential. ‘The new elite of union militants employed both old and new techniques of coercion to make workers labor harder and produce more’ (96). This led to manifold forms of resistance against work: absenteeism, fake illnesses, lateness, strikes, theft, sabotage, slowdowns, indiscipline and indifference (8). The CNT reacted with harsh punishments; their Minister of Justice, Juan García Oliver, even initiated ‘concentration camps’ with forced labour which were lauded by CNT militants as ‘progressive’ because they replaced torture and were designed to change the prisoner’s soul and values in a productivist way (99). But the workers themselves did not stop to resist work. Seidman speculates that ‘the way to eliminate resistance is not by workers’ control of the means of production but rather by the abolition of wage labor itself’ (17). Transformation of Work and Silent Resistance to Work So both Marxists and anarchists have a critical view of capitalism but are ambiguous about one of capitalism’s core features: work. At least, this was the case in the late nineteenth and early twentieth centuries. But work has changed since then, hasn’t it? The ‘scientific management’ of Frederick Taylor, initially adopted to increase productivity through decreasing worker’s shirking, has evolved into a largely automated and computerised industry – but a process of deskilling is still in force, stripping workers (even in the tertiary sector) of pride and dignity (Fineman, 2012: 56–57). Workers react to this by inventing their own work practices, sometimes by informal self-organisation (including punishing too productive co-workers), and sometimes just by bitter humour (58–59). Attempts by managers to use workers’ resistant tactics for profit optimisation – imposed ‘fun’, job ‘enrichment’ or forced ‘empowerment’ (59–62) – rarely have emancipatory outcomes. ‘For example, having stripped out layers of supervision to save costs, empowerment can appear as an excuse by management to intensify work, pushing more work and responsibility onto relatively fewer people, often for no extra reward’ (62). In spite of the health risks of long working hours, some workers adopt a ‘presenteeism’ – constant presence – for fear that not constantly being seen at the workplace risks career or job security (63–64). Others choose the opposite strategy: downshifting, i.e. switching to jobs that demand and pay less. This is ‘a radical response to the emptiness that some people feel as they chase ever more goods and services’ (66), but it suits only those with a certain level of wealth. If working less is involuntary, it is called ‘underemployment’ and often results in demoralisation, stress and illness (106). Similarly, unemployment does not free the worker from the necessity to work, only from the possibility. It is very widespread today and it serves as a threat to job holders. Today, you do not want a job because you like the job – you want it because joblessness is made highly inimical to good life. And nonetheless, people flee work in astonishing numbers, entailing much research on absenteeism. According to an official report, sickness absences cost the UK economy £15 billion in lost output, £9 billion in financial costs for the employers and £13 billion for the state – every year (Black and Frost, 2011: 14). This is the result of the 140 million working days lost per year in the UK. ‘This equates to 2.2 per cent of all working time, or 4.9 days for each worker each year, and is broadly comparable to many other developed countries (the United States, France, Germany and the Netherlands have similar rates)’ (19). This report ‘is premised on the fact that work is good for health in most cases’, at least when work is ‘good’ (18). Let’s put aside the incredibleness of this ‘fact’ for a moment and consider how these many sickness absences can be explained. One might think, after all, that if there was something wrong with the workplace, workers could use regular methods of industrial action to improve the situation without questioning the work regime as such. But a meta-analysis of eighty-two workplace ethnographies (covering 160 years in Britain and the US) showed that strikes occurred in approximately 21% of the cases, while absenteeism occurred in 45% of the cases (work avoidance at the workplace: 57%; social sabotage: 43%; playing dumb: 27%; theft: 20%) (Roscigno and Hodson, 2004: 22). While collective resistance (i.e. striking) is positively correlated with the presence of conflict and unions at the workplace (unions without conflict actually lessen the likelihood of strike action), individualised resistance like absenteeism does not have these requirements and can be used by workers in many circumstances (29, 31, 34). Because ‘collective and more individualized forms of resistance are not mutually exclusive in terms of their emergence or the factors that drive them’ (33), absenteeism is an obvious and commonplace tactic of class struggle that opens the door for anti-work action. Absenteeism is notoriously difficult to study because it is a ‘mildly deviant behavior’, so people will not always tell the truth to researchers (Johns, 2003: 159); for statistical analyses, ‘there are simply far too many intervening factors (known and unknown)’ (Hoxsey, 2010: 553). All findings cited below should therefore be read with caution – not least because their normative stance is to get as many people to work as much as possible. Shi and Skuterud (2015: 403) have shown that a rising recreational quality of the weather increases short-term sickness absence. Since good weather does not make people sick, it is plausible to assume that skipping workers can think of better things to do than work. By taking a sickie, workers express their preference for less hours of work (Livanos and Zangelidis, 2013: 494); it is their ‘labor-leisure choice’ (Lusinyan and Bonato, 2007: 477), made in a direct action rather than by pleading to superiors. It has to be noted, though, that this direct action happens in the framework of labour laws and often results in salary continuation, so it is not a valid indicator for deliberate anarchist attitudes. Although not enough attention has yet been given ‘to the meaning and role of absenteeism from the absentee’s perspective’, it can be said that there is at least ‘a modest link between absenteeism and job satisfaction’ (Hackett, 1989: 246). Workers often dislike their job and thus try to avoid it.

#### Pro-work arguments are self-interested propaganda by the rich for the poor to serve their masters – a post-work society intrinsically pleasure and work not as a means to a capitalistic end

Russell 32 [Bertrand Russell, philosopher, logician, and mathematician who was professor at the City College of New York, 1932, "In Praise of Idleness, by Bertrand Russell", Harper's Magazine, https://harpers.org/archive/1932/10/in-praise-of-idleness/]/Kankee

Throughout Europe, though not in America, there is a third class of men, more respected than either of the classes of workers. These are men who, through ownership of land, are able to make others pay for the privilege of being allowed to exist and to work. These landowners are idle, and I might, therefore, be expected to praise them. Unfortunately, their idleness is rendered possible only by the industry of others; indeed their desire for comfortable idleness is historically the source of the whole gospel of work. The last thing they have ever wished is that others should follow their example. From the beginning of civilization until the industrial revolution a man could, as a rule, produce by hard work little more than was required for the subsistence of himself and his family, although his wife worked at least as hard and his children added their labor as soon as they were old enough to do so. The small surplus above bare necessaries was not left to those who produced it, but was appropriated by priests and warriors. In times of famine there was no surplus; the warriors and priests, however, still secured as much as at other times, with the result that many of the workers died of hunger. This system persisted in Russia until 1917, and still persists in the East; in England, in spite of the Industrial Revolution, it remained in full force throughout the Napoleonic wars, and until a hundred years ago, when the new class of manufacturers acquired power. In America the system came to an end with the Revolution, except in the South, where it persisted until the Civil War. A system which lasted so long and ended so recently has naturally left a profound impression upon men’s thoughts and opinions. Much that we take for granted about the desirability of work is derived from this system and, being pre-industrial, is not adapted to the modern world. Modern technic has made it possible for leisure, within limits, to be not the prerogative of small privileged classes, but a right evenly distributed throughout the community. The morality of work is the morality of slaves, and the modern world has no need of slavery. It is obvious that, in primitive communities, peasants, left to themselves, would not have parted with the slender surplus upon which the warriors and priests subsisted, but would have either produced less or consumed more. At first sheer force compelled them to produce and part with the surplus. Gradually, however, it was found possible to induce many of them to accept an ethic according to which it was their duty to work hard, although part of their work went to support others in idleness. By this means the amount of compulsion required was lessened, and the expenses were diminished. To this day ninety-nine per cent of British wage-earners would be genuinely shocked if it were proposed that the King should not have a larger income than a working man. The conception of duty, speaking historically, has been a means used by the holders of power to induce others to live for the interests of their masters rather than their own. Of course the holders of power conceal this fact from themselves by managing to believe that their interests are identical with the larger interests of humanity. Sometimes this is true; Athenian slave-owners, for instance, employed part of their leisure in making a permanent contribution to civilization which would have been impossible under a just economic system. Leisure is essential to civilization, and in former times leisure for the few was rendered possible only by the labors of the many. But their labors were valuable, not because work is good, but because leisure is good. And with modern technic it would be possible to distribute leisure justly without injury to civilization. Modern technic has made it possible to diminish enormously the amount of labor necessary to produce the necessaries of life for every one. This was made obvious during the War. At that time all the men in the armed forces, all the men and women engaged in the production of munitions, all the men and women engaged in spying, war propaganda, or government offices connected with the War were withdrawn from productive occupations. In spite of this, the general level of physical well-being among wage-earners on the side of the Allies was higher than before or since. The significance of this fact was concealed by finance; borrowing made it appear as if the future was nourishing the present. But that, of course, would have been impossible; a man cannot eat a loaf of bread that does not yet exist. The War showed conclusively that by the scientific organization of production it is possible to keep modern populations in fair comfort on a small part of the working capacity of the modern world. If at the end of the War the scientific organization which had been created in order to liberate men for fighting and munition work had been preserved, and the hours of work had been cut down to four, all would have been well. Instead of that, the old chaos was restored, those whose work was demanded were made to work long hours, and the rest were left to starve as unemployed. Why? Because work is a duty, and a man should not receive wages in proportion to what he has produced, but in proportion to his virtue as exemplified by his industry. This is the morality of the Slave State, applied in circumstances totally unlike those in which it arose. No wonder the result has been disastrous. Let us take an illustration. Suppose that at a given moment a certain number of people are engaged in the manufacture of pins. They make as many pins as the world needs, working (say) eight hours a day. Someone makes an invention by which the same number of men can make twice as many pins as before. But the world does not need twice as many pins: pins are already so cheap that hardly any more will be bought at a lower price. In a sensible world everybody concerned in the manufacture of pins would take to working four hours instead of eight, and everything else would go on as before. But in the actual world this would be thought demoralizing. The men still work eight hours, there are too many pins, some employers go bankrupt, and half the men previously concerned in making pins are thrown out of work. There is, in the end, just as much leisure as on the other plan, but half the men are totally idle while half are still overworked. In this way it is insured that the unavoidable leisure shall cause misery all round instead of being a universal source of happiness. Can anything more insane be imagined? The idea that the poor should have leisure has always been shocking to the rich. In England in the early nineteenth century fifteen hours was the ordinary day’s work for a man; children sometimes did as much, and very commonly did twelve hours a day. When meddlesome busy-bodies suggested that perhaps these hours were rather long, they were told that work kept adults from drink and children from mischief. When I was a child, shortly after urban working men had acquired the vote, certain public holidays were established by law, to the great indignation of the upper classes. I remember hearing an old Duchess say, “What do the poor want with holidays? they ought to work.” People nowadays are less frank, but the sentiment persists, and is the source of much economic confusion. II Let us, for a moment, consider the ethics of work frankly, without superstition. Every human being, of necessity, consumes in the course of his life a certain amount of produce of human labor. Assuming, as we may, that labor is on the whole disagreeable, it is unjust that a man should consume more than he produces. Of course he may provide services rather than commodities, like a medical man, for example; but he should provide something in return for his board and lodging. To this extent, the duty of work must be admitted, but to this extent only. I shall not develop the fact that in all modern societies outside the U. S. S. R. many people escape even this minimum of work, namely all those who inherit money and all those who marry money. I do not think the fact that these people are allowed to be idle is nearly so harmful as the fact that wage-earners are expected to overwork or starve. If the ordinary wage-earner worked four hours a day there would be enough for everybody, and no unemployment — assuming a certain very moderate amount of sensible organization. This idea shocks the well-to-do, because they are convinced that the poor would not know how to use so much leisure. In America men often work long hours even when they are already well-off; such men, naturally, are indignant at the idea of leisure for wage-earners except as the grim punishment of unemployment, in fact, they dislike leisure even for their sons. Oddly enough, while they wish their sons to work so hard as to have no time to be civilized, they do not mind their wives and daughters having no work at all. The snobbish admiration of uselessness, which, in an aristocratic society, extends to both sexes, is under a plutocracy confined to women; this, however, does not make it any more in agreement with common sense. The wise use of leisure, it must be conceded, is a product of civilization and education. A man who has worked long hours all his life will be bored if he becomes suddenly idle. But without a considerable amount of leisure a man is cut off from many of the best things. There is no longer any reason why the bulk of the population should suffer this deprivation; only a foolish asceticism, usually vicarious, makes us insist on work in excessive quantities now that the need no longer exists. In the new creed which controls the government of Russia, while there is much that is very different from the traditional teaching of the West, there are some things that are quite unchanged. The attitude of the governing classes, and especially of those who control educational propaganda, on the subject of the dignity of labor is almost exactly that which the governing classes of the world have always preached to what were called the “honest poor.” Industry, sobriety, willingness to work long hours for distant advantages, even submissiveness to authority, all these reappear; moreover, authority still represents the will of the Ruler of the Universe, Who, however, is now called by a new name, Dialectical Materialism. The victory of the proletariat in Russia has some points in common with the victory of the feminists in some other countries. For ages men had conceded the superior saintliness of women and had consoled women for their inferiority by maintaining that saintliness is more desirable than power. At last the feminists decided that they would have both, since the pioneers among them believed all that the men had told them about the desirability of virtue but not what they had told them about the worthlessness of political power. A similar thing has happened in Russia as regards manual work. For ages the rich and their sycophants have written in praise of “honest toil,” have praised the simple life, have professed a religion which teaches that the poor are much more likely to go to heaven than the rich, and in general have tried to make manual workers believe that there is some special nobility about altering the position of matter in space, just as men tried to make women believe that they derived some special nobility from their sexual enslavement. In Russia all this teaching about the excellence of manual work has been taken seriously, with the result that the manual worker is more honored than anyone else. What are, in essence, revivalist appeals are made to secure shock workers for special tasks. Manual work is the ideal which is held before the young, and is the basis of all ethical teaching. For the present this is all to the good. A large country, full of natural resources, awaits development and has to be developed with very little use of credit. In these circumstances hard work is necessary and is likely to bring a great reward. But what will happen when the point has been reached where everybody could be comfortable without working long hours? In the West we have various ways of dealing with this problem. We have no attempt at economic justice, so that a large proportion of the total produce goes to a small minority of the population, many of whom do no work at all. Owing to the absence of any central control over production, we produce hosts of things that are not wanted. We keep a large percentage of the working population idle because we can dispense with their labor by making others overwork. When all these methods prove inadequate we have a war: we cause a number of people to manufacture high explosives, and a number of others to explode them, as if we were children who had just discovered fireworks. By a combination of all these devices we manage, though with difficulty, to keep alive the notion that a great deal of manual work must be the lot of the average man. In Russia, owing to economic justice and central control over production, the problem will have to be differently solved. The rational solution would be as soon as the necessaries and elementary comforts can be provided for all to reduce the hours of labor gradually, allowing a popular vote to decide, at each stage, whether more leisure or more goods were to be preferred. But, having taught the supreme virtue of hard work, it is difficult to see how the authorities can aim at a paradise in which there will be much leisure and little work. It seems more likely that they will find continually fresh schemes by which present leisure is to be sacrificed to future productivity. I read recently of an ingenious scheme put forward by Russian engineers for making the White Sea and the northern coasts of Siberia warm by putting a dam across the Kara Straits. An admirable plan, but liable to postpone proletarian comfort for a generation, while the nobility of toil is being displayed amid the ice-fields and snowstorms of the Arctic Ocean. This sort of thing, if it happens, will be the result of regarding the virtue of hard work as an end in itself, rather than as a means to a state of affairs in which it is no longer needed. III The fact is that moving matter about, while a certain amount of it is necessary to our existence, is emphatically not one of the ends of human life. If it were, we should have to consider every navvy superior to Shakespeare. We have been misled in this matter by two causes. One is the necessity of keeping the poor contented, which has led the rich for thousands of years to preach the dignity of labor, while taking care themselves to remain undignified in this respect. The other is the new pleasure in mechanism, which makes us delight in the astonishingly clever changes that we can produce on the earth’s surface. Neither of these motives makes any great appeal to the actual worker. If you ask him what he thinks the best part of his life, he is not likely to say, “I enjoy manual work because it makes me feel that I am fulfilling man’s noblest task, and because I like to think how much man can transform his planet. It is true that my body demands periods of rest, which I have to fill in as best I may, but I am never so happy as when the morning comes and I can return to the toil from which my contentment springs.” I have never heard working men say this sort of thing. They consider work, as it should be considered, as a necessary means to a livelihood, and it is from their leisure hours that they derive whatever happiness they may enjoy. It will be said that while a little leisure is pleasant, men would not know how to fill their days if they had only four hours’ work out of the twenty-four. In so far as this is true in the modern world it is a condemnation of our civilization; it would not have been true at any earlier period. There was formerly a capacity for light-heartedness and play which has been to some extent inhibited by the cult of efficiency. The modern man thinks that everything ought to be done for the sake of something else, and never for its own sake. Serious-minded persons, for example, are continually condemning the habit of going to the cinema, and telling us that it leads the young into crime. But all the work that goes to producing a cinema is respectable, because it is work, and because it brings a money profit. The notion that the desirable activities are those that bring a profit has made everything topsy-turvy. The butcher who provides you with meat and the baker who provides you with bread are praiseworthy because they are making money but when you enjoy the food they have provided you are merely frivolous, unless you eat only to get strength for your work. Broadly speaking, it is held that getting money is good and spending money is bad. Seeing that they are two sides of one transaction, this is absurd; one might as well maintain that keys are good but keyholes are bad. The individual, in our society, works for profit; but the social purpose of his work lies in the consumption of what he produces. It is this divorce between the individual and the social purpose of production that makes it so difficult for men to think clearly in a world in which profitmaking is the incentive to industry. We think too much of production and too little of consumption. One result is that we attach too little importance to enjoyment and simple happiness, and that we do not judge production by the pleasure that it gives to the consumer. When I suggest that working hours should be reduced to four, I am not meaning to imply that all the remaining time should necessarily be spent in pure frivolity. I mean that four hours’ work a day should entitle a man to the necessities and elementary comforts of life, and that the rest of his time should be his to use as he might see fit. It is an essential part of any such social system that education should be carried farther than it usually is at present, and should aim, in part, at providing tastes which would enable a man to use leisure intelligently. I am not thinking mainly of the sort of things that would be considered “high-brow.” Peasant dances have died out except in remote rural areas, but the impulses which caused them to be cultivated must still exist in human nature. The pleasures of urban populations have become mainly passive: seeing cinemas, watching football matches, listening to the radio, and so on. This results from the fact that their active energies are fully taken up with work; if they had more leisure they would again enjoy pleasures in which they took an active part. In the past there was a small leisure class and a large working class. The leisure class enjoyed advantages for which there was no basis in social justice; this necessarily made it oppressive, limited its sympathies, and caused it to invent theories by which to justify its privileges. These facts greatly diminished its excellence, but in spite of this drawback it contributed nearly the whole of what we call civilization. It cultivated the arts and discovered the sciences; it wrote the books, invented the philosophies, and refined social relations. Even the liberation of the oppressed has usually been inaugurated from above. Without the leisure class mankind would never have emerged from barbarism. The method of a hereditary leisure class without duties was, however, extraordinarily wasteful. None of the members of the class had been taught to be industrious, and the class as a whole was not exceptionally intelligent. It might produce one Darwin, but against him had to be set tens of thousands of country gentlemen who never thought of anything more intelligent than fox-hunting and punishing poachers. At present, the universities are supposed to provide, in a more systematic way, what the leisure class provided accidentally and as a byproduct. This is a great improvement, but it has certain drawbacks. University life is so different from life in the world at large that men who live in an academic milieu tend to be unaware of the pre-occupations of ordinary men and women; moreover, their ways of expressing themselves are usually such as to rob their opinions of the influence that they ought to have upon the general public. Another disadvantage is that in universities studies are organized, and the man who thinks of some original line of research is likely to be discouraged. Academic institutions, therefore, useful as they are, are not adequate guardians of the interests of civilization in a world where every one outside their walls is too busy for unutilitarian pursuits. In a world where no one is compelled to work more than four hours a day every person possessed of scientific curiosity will be able to indulge it, and every painter will be able to paint without starving, however excellent his pictures may be. Young writers will not be obliged to draw attention to themselves by sensational pot-boilers, with a view to acquiring the economic independence needed for monumental works, for which, when the time at last comes, they will have lost the taste and the capacity. Men who in their professional work have become interested in some phase of economics or government will be able to develop their ideas without the academic detachment that makes the work of university economists lacking in reality. Medical men will have time to learn about the progress of medicine. Teachers will not be exasperatedly struggling to teach by routine things which they learned in their youth, which may, in the interval, have been proved to be untrue. Above all, there will be happiness and joy of life, instead of frayed nerves, weariness, and dyspepsia. The work exacted will be enough to make leisure delightful, but not enough to produce exhaustion. Since men will not be tired in their spare time, they will not demand only such amusements as are passive and vapid. At least one per cent will probably devote the time not spent in professional work to pursuits of some public importance, and, since they will not depend upon these pursuits for their livelihood, their originality will be unhampered, and there will be no need to conform to the standards set by elderly pundits. But it is not only in these exceptional cases that the advantages of leisure will appear. Ordinary men and women, having the opportunity of a happy life, will become more kindly and less persecuting and less inclined to view others with suspicion. The taste for war will die out, partly for this reason, and partly because it will involve long and severe work for all. Good nature is, of all moral qualities, the one that the world needs most, and good nature is the result of ease and security, not of a life of arduous struggle. Modern methods of production have given us the possibility of ease and security for all; we have chosen instead to have overwork for some and starvation for others. Hitherto we have continued to be as energetic as we were before there were machines. In this we have been foolish, but there is no reason to go on being foolish for ever.

#### AI post-work allows freedom and time for voluntary labor and escape from workplace stress

Garimella 18 [Kiran Garimella, Assistant Professor in the School of Communication and Information at Rutgers University, 8-7-2018, "Job Loss From AI? There's More To Fear!", Forbes, https://www.forbes.com/sites/cognitiveworld/2018/08/07/job-loss-from-ai-theres-more-to-fear/]/Kankee

How would you adapt to a revolution that promises, not in words but by example, to eliminate the notion of jobs itself? This is the unthinkable frontier beyond which few people are venturing. The idea that we must have jobs is so ingrained in human culture that we forget to ask why. Let's step back for a second and look at this objectively. The reason we need jobs is so we can pay our bills, buy things, and fulfill our desires. What we really want aren't jobs but the income that jobs bring us. What we need is work (not jobs) for self-fulfillment, and therein lies an enormous difference. More on that below. What proportion of workers can truly claim that they love their jobs? Talk is cheap. The acid test is this: If you were to win a lottery jackpot worth several millions, would you continue to work at your current job, perhaps for free? I suspect that even those who claim that they really, really love their jobs would re-examine their options if they hit the jackpot. People may continue working, but they wouldn't hanker after paying jobs, especially not jobs that are ridden with stress, abuse, harassment, and office politics. Isn't this the promise of automation? Let's take a walk down a path that admittedly stretches our credulity. Imagine a future where most of the work is automated, thanks to AI. This would include farming, transportation of goods, personal transportation, maintenance of roads, generation of power, manufacturing, custodial services, interaction services (for example, customer service and in restaurants), medical care, elderly care, schooling, and so on. Note that I'm not saying humans will be—or even could be—completely eliminated from the equation, only that their role will become either minimal or move up the cognitive scale; they would function as casual supervisors or auxiliary helpers when things get really tricky, and that too on very reduced work hours. Far out? Not at all. This is already happening in isolated examples. As William Gibson said, "The future is already here - it's just not very evenly distributed." Informed people know that this is already happening. What stretches the imagination a bit is how this affects jobs. Let's get back to the fundamentals of economics - the law of supply and demand and its effect on prices. When AI eliminates jobs (more accurately, the need for them), there is the obvious loss of income. This means less disposable income and reduction of spending on nice-to-have goods and luxuries. Less demand forces prices to drop. If prices drop below a level where commodity margins are threatened, the company and eventually the industry, will fold. Prices of necessary goods (generally, commodities) will continue to drop, but shrinking margins are somewhat offset by decreasing operational costs (thanks to AI-driven automation). Food prices, for example, could go down. Remember that the sinister food cartels that we fear (whether truly sinister or not) cannot keep charging seemingly high prices because consumers' inability to pay cannot sustain high prices in the long run. Rather than outright and widespread job loss, what I think will happen is a gradual lessening of work hours, eventually moving employees from full-time to temp. But, if all things go well, the loss in pay need not pinch because there will be a corresponding decrease in costs for all types of goods. People will slowly come to the realization that total income is less important than relative income. Indeed, income that is measured relative to one's neighbor is even less important than measuring against its economic power. Would you lose much sleep if your six-figure salary got cut in half if home prices were also slashed to a fifth, a pound of filet mignon cost $1.50, and a 10-lb bag of potatoes cost 10 cents? Would you really care if the richest person on the planet makes one hundred million times more income than you do under these conditions? I don't think people would worry so much about the end result if all things go well. The problem is in the transition, of course. Not all things go well all the time. But the general direction will be towards less job-related work without any corresponding reduction in the quality of life (and perhaps an overall improvement in the quality of lifestyle). This progress will obviously not happen in a straight, happy path, but will meander agonizingly, causing considerable distress, backtracking, and excitement along the way. In truth, I suspect that there will be plenty of advance notice of impending job loss. A few hundred thousand truck drivers aren't going to show up for work one morning only to be told that they have all been replaced overnight by driverless trucks. Drivers' ed instructors will not be completely blindsided to find one day that they have no students because everyone switched over to driverless cars. The entire DMV staff isn't going to be laid off on the same day because autonomous vehicles and blockchains put them out of business overnight. There will be a transition which, despite all the hype and exuberance, will have a fairly long trajectory. But slowly, slowly, across the board, people will start discovering that they really don't have to work 60 hours a week. There will be serious proposals to reduce the work week from 40 hours to 36 hours; then to four days a week. Full-time work gives way to part-time work. Desk-bound, 8-5 work will slowly give way to flex-time work. Remote work will edge out on-site work. People will cut their working hours by 25% and find, to their pleasant surprise, that their lifestyle actually improves a bit as a result, even if they aren't making the previous level of income. Hard work will be replaced by smart work (but 'smart' perhaps not in the sense of making enormous amounts of money with little work, but in the sense of obtaining high returns on happiness and lifestyle). Financial freedom will give way to time freedom. AI will cause enormous job losses, but that is nothing compared to the loss of meaning in traditional work and virtue of duty-bound toil. This will force people to re-examine their motivations for work. Would you get up in the morning and do something for your own sake, for your own pleasure and passion, when you don't have to do so to feed yourself? What will you do when you don't have to? Many people will be unpleasantly shocked to find they have no passion. For them, AI spells debilitating inertia, deepening despair, and the slide into apathy. A happy group (hopefully, more than a few) will look upon this as a golden opportunity to get rid of the shackles of work they must do so they can do work they want to do. The AI revolution will be eclipsed by the revolution caused by decoupling jobs from work and by moving from "I work, therefore I live" to "I live, therefore I work." Sociologists will be forced to re-examine and re-formulate their models of human interaction and organization. Economists will be forced to re-think incentives and agency relationships. Politicians will be forced to invent new rhetoric for their platforms when the traditional political posturings will become moot. Schools will be forced to contend with a deschooled society (in the sense of Ivan Illich's "Deschooling Society"). Society as a whole must then grapple with the deeper social, economic, and psychological ramifications of permanent net job losses caused by AI. In consolation, the loss of jobs without (optimistically) loss of lifestyle should give us the time and freedom to think about these issues. Perhaps we can even debate the nature of intelligence as it applies to artificial intelligence. If all this pans out as I think it will and should, this is quite a good impact of artificial intelligence on natural intelligence; we have nothing to fear but to be prepared for the transition itself!

#### Mass leisure/UBI require AGI to be economically feasible

Pethokoukis 24 [James Pethokoukis, Senior Fellow DeWitt Wallace Chair Editor and a graduate of Northwestern University and the Medill School of Journalism, 8-9-2024, "UBI Needs AGI to Make Sense", American Enterprise Institute, https://www.aei.org/articles/ubi-needs-agi-to-make-sense/]/Kankee

The case for universal basic income is in large part aesthetic. The minimalist simplicity and elegance of UBI constitute a big part of its appeal: a single solution addressing a multitude of societal complexities. By providing all citizens an unconditional stipend, UBI tackles poverty, inequality, and economic instability — especially from technological flux — in one swooping, graceful motion. Its streamlined structure bypasses bureaucratic bottlenecks and tangles, while its inherent adaptability allows individuals to create their own life paths. Just lovely. Public policy by way of Jony Ive. Imagine someone in Silicon Valley who worries about tech unemployment from human-level artificial general intelligence. Also imagine such a person stumbling upon UBI after a quick Google query and thinking, “Money for people. What an obvious solution. I must tell the world of this, especially those gloomy idiots in Washington.” But here’s the thing: America already has a rather extensive welfare state, and it’s deeply embedded in our society. So many programs: Social Security to provide retirement, disability, and survivor benefits. Medicare to offer health insurance for seniors and certain disabled individuals. Medicaid to supply health coverage for low-income populations. SNAP to furnish food assistance to low-income individuals and families. TANF to deliver cash aid and support services to needy families with children. SSI to grant cash assistance to low-income aged, blind, or disabled persons. Section 8 to facilitate affordable housing for low-income families. The Earned Income Tax Credit to extend tax credits to low- and moderate-income workers. The Child Tax Credit to afford financial support to families with children. Unemployment Insurance to dispense temporary aid to eligible jobless workers. So, yeah, it’s a lot. But it also does a lot. (And my brilliant AEI colleague spend a lot of their time figuring out how to make it all work even better and get more bang for the buck.) Consider: From 1979 through 2022, middle-class income (inflation-adjusted) rose by 43 percent, according to a Congressional Budget Office calculation that includes wage, business, and investment income, as well as payments from government programs that people pay into while working, such as Social Security and Medicare. That number jumps to 76 percent if you include means-tested transfers — Medicaid and children’s health insurance, SNAP, Supplemental Security Income, housing assistance — and refundable tax credits. No doubt the idea of swapping our tangled welfare system for a simple and elegant UBI is appealing in theory. But as with totally changing the US tax code, there needs to be considerable and demonstrable upside to seriously consider such a heavy lift. And I interpret the new research on UBI, supported by OpenAI’s Sam Altman, as failing to provide such compelling up-side evidence. The project tracked 1,000 low-income individuals given $1,000 monthly for three years. Compared to a control group, UBI recipients worked less, earned less, and enjoyed more leisure time. Their income fell by about $1,500 — not counting the UBI payments — annually due to reduced work hours and labor participation. For every $1 received, household income dropped by over 20 cents. (“This is a pretty substantial effect.” ) The extra time wasn’t used for job searching or entrepreneurship, but mainly for leisure activities. Younger participants were slightly more likely to pursue education. While the study found “precursors” to entrepreneurialism — “entrepreneurial orientation and intention” — it didn’t show significant increases in business creation or human capital investments. (“In short: the transfers largely financed consumption, with minimal savings and ~$0 impact on net worth. Financial health improved in years 1-2 but reverted by year 3.”) Nor did physical health and mental health improve, really. (There was no effect of the transfer on physical health, measured via self-reports, clinical outcomes derived from blood draws, and admin records of mortality.” … “The cash generated big improvements in stress and mental health, but they were short-lived.”) This from my AEI colleague Kevin Corinth: The employment declines in these studies are important, despite the fact that they are based on short-term cash infusions that likely understate the effect of a permanent, nationwide policy. They suggest that a true Universal Basic Income (UBI), layered on top of existing benefits to low-income families, would reduce work effort. But they are only the tip of the iceberg when it comes to employment concerns. In reality, a UBI of $12,000 for every adult in the United States would be impossibly expensive, costing over $3 trillion per year. It is hard to imagine Congress greenlighting a policy to increase federal spending by more than 50 percent—not to mention reigniting inflation—while already-high federal debt continues to climb. As Corinth sees it, a more constructive approach to welfare reform would balance immediate support with long-term self-sufficiency. This involves offering temporary aid during hardships, expanding work requirements for able-bodied adults receiving long-term assistance, and reducing marriage penalties in the system. And as for replacing the entire safety net with a UBI — setting aside political feasibility — the odds seem high that the simple and elegant system wouldn’t remain that way for long. My AEI colleague Michael Strain writes: Another issue with UBI is its lack of realism. If UBI were introduced here, it wouldn’t take long for a politician to point out that, say, blind people need more support than those without physical disabilities. And then that workers who are disabled on the job deserve extra support over and above their base-line UBI benefit. And then it wouldn’t take very long for UBI to transform into something that looks very much like the system we have today. Why, then, change in the first place? As I see it, the best case for a UBI remains a science-fictional one: a world where labor markets are completely upended by brilliant machines able to match or exceed humans in both cognitive and physical capabilities. So not a today thing, though certainly worth continuing to study and discuss. While AGI’s timeline remains uncertain, let’s consider a scenario where AI significantly reduces workforce participation but there are still lots of jobs for humans. Rather than focusing solely on distributing AI-driven income, governments should prioritize making work as financially attractive as possible and helping people leverage AI to boost their productivity. This could involve substantial earnings subsidies, with many workers possibly receiving most of their income from government checks rather than employers. Such an approach, while not ideal, would be preferable to a world where these individuals contribute nothing and rely entirely on smart machine-generated wealth.

#### AGI allows techno-communism and a post-work future

Xiang 18 [Feng Xiang, legal scholar at Tsinghua University, 5-3-2018, "AI will spell the end of capitalism", Washington Post, https://www.washingtonpost.com/news/theworldpost/wp/2018/05/03/end-of-capitalism/]/Kankee

The most momentous challenge facing socio-economic systems today is the arrival of artificial intelligence. If AI remains under the control of market forces, it will inexorably result in a super-rich oligopoly of data billionaires who reap the wealth created by robots that displace human labor, leaving massive unemployment in their wake. But China’s socialist market economy could provide a solution to this. If AI rationally allocates resources through big data analysis, and if robust feedback loops can supplant the imperfections of “the invisible hand” while fairly sharing the vast wealth it creates, a planned economy that actually works could at last be achievable. The more AI advances into a general-purpose technology that permeates every corner of life, the less sense it makes to allow it to remain in private hands that serve the interests of the few instead of the many. More than anything else, the inevitability of mass unemployment and the demand for universal welfare will drive the idea of socializing or nationalizing AI. Marx’s dictum, “From each according to their abilities, to each according to their needs,” needs an update for the 21st century: “From the inability of an AI economy to provide jobs and a living wage for all, to each according to their needs.” Even at this early stage, the idea that digital capitalism will somehow make social welfare a priority has already proven to be a fairytale. The billionaires of Google and Apple, who have been depositing company profits in offshore havens to avoid taxation, are hardly paragons of social responsibility. The ongoing scandal around Facebook’s business model, which puts profitability above responsible citizenship, is yet another example of how in digital capitalism, private companies only look after their own interests at the expense of the rest of society. One can readily see where this is all headed once technological unemployment accelerates. “Our responsibility is to our shareholders,” the robot owners will say. “We are not an employment agency or a charity.” These companies have been able to get away with their social irresponsibility because the legal system and its loopholes in the West are geared to protect private property above all else. Of course, in China, we have big privately owned Internet companies like Alibaba and Tencent. But unlike in the West, they are monitored by the state and do not regard themselves as above or beyond social control. It is the very pervasiveness of AI that will spell the end of market dominance. The market may reasonably if unequally function if industry creates employment opportunities for most people. But when industry only produces joblessness, as robots take over more and more, there is no good alternative but for the state to step in. As AI invades economic and social life, all private law-related issues will soon become public ones. More and more, regulation of private companies will become a necessity to maintain some semblance of stability in societies roiled by constant innovation. I consider this historical process a step closer to a planned market economy. Laissez-faire capitalism as we have known it can lead nowhere but to a dictatorship of AI oligarchs who gather rents because the intellectual property they own rules over the means of production. On a global scale, it is easy to envision this unleashed digital capitalism leading to a battle between robots for market share that will surely end as disastrously as the imperialist wars did in an earlier era. For the sake of social well-being and security, individuals and private companies should not be allowed to possess any exclusive cutting-edge technology or core AI platforms. Like nuclear and biochemical weapons, as long as they exist, nothing other than a strong and stable state can ensure society’s safety. If we don’t nationalize AI, we could sink into a dystopia reminiscent of the early misery of industrialization, with its satanic mills and street urchins scrounging for a crust of bread. The dream of communism is the elimination of wage labor. If AI is bound to serve society instead of private capitalists, it promises to do so by freeing an overwhelming majority from such drudgery while creating wealth to sustain all. If the state controls the market, instead of digital capitalism controlling the state, true communist aspirations will be achievable. And because AI increasingly enables the management of complex systems by processing massive amounts of information through intensive feedback loops, it presents, for the first time, a real alternative to the market signals that have long justified laissez-faire ideology — and all the ills that go with it. Going forward, China’s socialist market economy, which aims to harness the fruits of production for the whole population and not just a sliver of elites operating in their own self-centered interests, can lead the way toward this new stage of human development. If properly regulated in this way, we should celebrate, not fear, the advent of AI. If it is brought under social control, it will finally free workers from peddling their time and sweat only to enrich those at the top. The communism of the future ought to adopt a new slogan: “Robots of the world, unite!”

### Contention 6: AGI Securitization

#### AI threat construction threatens purposeful nuclear war, domestic terrorism, and genocide

Torres 23 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 8-23-2023, "A Code Red Warning about TESCREALism", Truthdig, https://www.truthdig.com/articles/before-its-too-late-buddy/]/Kankee

In 2023, these worries are no longer hypothetical. First, the TESCREAL movement has become immensely powerful. It’s infiltrating foreign policy circles and major governing institutions like the United Nations, has tens of billions of dollars behind it, is pervasive within Silicon Valley, and has been promoted by people with large social media followings. Elon Musk, for example, calls longtermism “a close match for my philosophy,” and last year retweeted a link to Bostrom’s paper about how many digital people there could be if we colonize space, along with the line: “Likely the most important paper ever written.” The TESCREAL movement has become a global force, and it’s building momentum. Second, some TESCREALists have begun to explicitly call for policies that could heighten the risk of nuclear conflicts that kill billions of people. Others have flirted with the idea of targeted assassinations of AI researchers to slow down progress on artificial general intelligence, or AGI, which many TESCREALists see as the greatest existential risk facing humanity this century. Talk of extreme actions, even the use of force and violence, to prevent an AGI apocalypse is becoming increasingly common, and my worry now is that true believers in the TESCREAL ideologies, who think we’re in an apocalyptic moment with AGI, could actually do something that causes serious harm to others. Consider a recent TIME magazine article by Eliezer Yudkowsky, a central figure within the TESCREAL movement who calls himself a “genius” and has built a cult-like following in the San Francisco Bay Area. Yudkowsky contends that we may be on the cusp of creating AGI, and that if we do this “under anything remotely like the current circumstances,” the “most likely result” will be “that literally everyone on Earth will die.” Since an all-out thermonuclear war probably won’t kill everyone on Earth—the science backs this up—he thus argues that countries should sign an international treaty that would sanction military strikes against countries that might be developing AGI, even at the risk of triggering a “full nuclear exchange.” Many people found these claims shocking. Three days after the article was published, someone asked Yudkowsky on social media: “How many people are allowed to die to prevent AGI?” His response was: “There should be enough survivors on Earth in close contact to form a viable reproductive population, with room to spare, and they should have a sustainable food supply. So long as that’s true, there’s still a chance of reaching the stars someday.” To understand just how extreme this is, a viable reproductive population might be as low as 150 people, although more conservative estimates put the number at 40,000. As of this writing, the current human population is approximately 8,054,000,000. If you subtract 40,000 from this number, you get 8,053,960,000. Yudkowsky is thus arguing that more than eight billion people should be “allowed” to die for the sake of “reaching the stars someday,” i.e., realizing the techno-utopian vision at the heart of TESCREALism. Astonishingly, after Yudkowsky published his article and made the comments above, TED invited him to give a talk. He also appeared on major podcast’s like Lex Fridman’s, and last month appeared on the “Hold These Truths Podcast” hosted by the Republican congressman Dan Crenshaw. The extremism that Yudkowsky represents is starting to circulate within the public and political arenas, and his prophecies about an imminent AGI apocalypse are gaining traction. The first time I became worried about what TESCREALism might “justify” was in 2018, when I still considered myself to be part of the movement. I read a book titled Here Be Dragons by the Swedish scholar Olle Häggström, who is generally sympathetic with the longtermist ideology. In one chapter, Häggström considers Bostrom’s claim that teeny-tiny reductions in existential risk could be orders of magnitude better than saving billions of actual human lives. He then writes that: I feel extremely uneasy about the prospect that [Bostrom’s calculations] might become recognised among politicians and decision-makers as a guide to policy worth taking literally. It is simply too reminiscent of the old saying “If you want to make an omelette, you must be willing to break a few eggs,” which has typically been used to explain that a bit of genocide or so might be a good thing, if it can contribute to the goal of creating a future utopia. Imagine a real-world scenario, Häggström says, in which the CIA explains to the US president that they have credible evidence that somewhere in Germany, there is a lunatic who is working on a doomsday weapon and intends to use it to wipe out humanity, and that this lunatic has a one-in-a-million chance of succeeding. They have no further information on the identity or whereabouts of this lunatic. If the president has taken Bostrom’s argument to heart, and if he knows how to do the arithmetic, he may conclude that it is worthwhile conducting a full-scale nuclear assault on Germany to kill every single person within its borders. When Häggström wrote this, it was just another hypothetical concern. Yet today, this scenario is eerily similar to what Yudkowsky is advocating: military actions that could cause a genocidal nuclear catastrophe, if necessary to keep the techno-utopian dream alive. Yudkowsky is careful to note that he doesn’t say countries should engage in nuclear first strikes, only conventional first strikes, though what does this matter if the result is a nuclear war that kills more than 8 billion people? As I have elsewhere argued, any time an ideology or worldview combines a utopian vision of the future marked by near-infinite value with a broadly “utilitarian” mode of moral reasoning—as TESCREALism does—it could easily lead true believers to conclude that, as Häggström writes, “a bit of genocide or so might be a good thing, if it can contribute to the goal of creating a future utopia.” This is a Code Red Warning about TESCREALism. If Yudkowsky’s words are taken seriously by our political leaders, or perhaps some lone wolves, we should be extremely worried about the harms that could result. Over and over again, history shows that the march to utopia can leave a trail of destruction in its wake. If the ends can justify the means, and the end is paradise, then what exactly is off the table for protecting and preserving this end? There are other worrying signs within the community as well. For example, last February someone sent me the meeting minutes of an invite-only workshop on AI safety. The field of “AI safety” emerged out of the TESCREAL movement. It’s based on the assumption that if we create an AGI that’s “aligned” with our “values,” it will immediately grant all our techno-utopian wishes, including immortality for everyone who wants it. However, if the AGI is “misaligned,” it will kill every human on the planet by default. The goal of AI safety research is to ensure that AGI is sufficiently aligned, although people like Yudkowsky have become convinced that we’re so far from solving the “alignment” problem that the only option left is to ban all AGI research—including through force. The AI safety workshop was held in Berkeley in late 2022 and initially funded by the FTX Future Fund, established by Sam Bankman-Fried, the TESCREAList who appears to have committed “one of the biggest financial frauds in American history.” Berkeley is the home town of Yudkowsky’s own Machine Intelligence Research Institute (MIRI), where one of the workshop organizers was subsequently employed. Although MIRI was not directly involved in the workshop, Yudkowsky reportedly attended a workshop afterparty. Under the heading “produce aligned AI before unaligned [AI] kills everyone,” the meeting minutes indicate that someone suggested the following: “Solution: be Ted Kaczynski.” Later on, someone proposed the “strategy” of “start building bombs from your cabin in Montana,” where Kaczynski conducted his campaign of domestic terrorism, “and mail them to DeepMind and OpenAI lol.” This was followed a few sentences later by, “Strategy: We kill all AI researchers.” Participants noted that if such proposals were enacted, they could be “harmful for AI governance,” presumably because of the reputational damage they might cause to the AI safety community. But they also implied that if all AI researchers are killed, this could mean that AGI doesn’t get built. And foregoing AGI, if properly aligned, would mean that we “lose a lot of potential value of good things.” It’s not clear how serious the participants were. But their discussion indicates that talk of violence is becoming normalized. Yudkowsky is partly to blame, given his cult-like status among many AI safety researchers fretting about the apocalypse. Not only has he made the claims quoted above, but he’s endorsed property damage targeting AI companies; and while he insists that “I have at no point uttered the words ‘b\*mb datacenters’ nor called for individual terrorism,” when someone asked him whether he would “have supported bombing the Wuhan center studying pathogens in 2019,” he said this: Great question! I’m at roughly 50% that they killed a few million people and cost many trillions of dollars [a reference to the lab leak theory]. If I can do it secretly, I probably do and then throw up a lot. If people are going to see it, it’s not worth the credibility hit on AGI, since nobody would know why I was doing that. I can definitely think of better things to do with the hypothetical time machine. Yudkowsky talks out of both sides of his mouth, giving mixed messages. He wants us to believe that his proposals for avoiding an AGI apocalypse are within the bounds of established norms, yet in a moment of honesty he says that he would “probably” bomb laboratories in Wuhan, if he can do this “secretly.” He opposes nuclear first strikes, yet implies that more than 8 billion people should be “allowed” to die, if it means keeping the door open to “reaching the stars someday.” And his extraordinary over-confidence, fueled by his egomania, that a misaligned AGI will kill everyone on Earth is inspiring the sort of radical, dangerous discussions like those recorded in the Berkeley workshop meeting minutes. It’s not just AI researchers who have been singled out: so have critics of the TESCREAL ideology. In fact, the reason the whistleblower leaked the meeting minutes was because he saw that I’ve received numerous threats of physical violence over the past year for speaking out against the movement I once belonged to. For example, in October of last year, I received an anonymous DM over Twitter saying: “Better be careful or an EA superhero will break your kneecaps.” This message was repeated the following day by a different Twitter account, which later wrote “I wish you remained unborn” under a social media post of mine. The next month, an anonymous person in the TESCREAL movement sent me a menacing email that read: “Get psychiatric assistance before it’s too late, buddy,” after which yet another anonymous account threatened to try to dox me. Last June, I received an email from an unidentified TESCREAList referring to a short film about a murder-suicide, in which a mother kills herself and her disabled daughter by setting fire to their car while both are inside. The sender stated, “I hope it will take something far less extreme than what happens in the film to make you look at the kind of person you’re becoming,” by which they meant a vigorous opponent of TESCREALism. I’m not the only one who’s been frightened by the TESCREAL community. Another critic of longtermism, Simon Knutsson, wrote in 2019 that he had become concerned about his safety, adding that he’s “most concerned about someone who finds it extremely important that there will be vast amounts of positive value in the future and who believes I stand in the way of that,” a reference to the TESCREAL vision of astronomical future value. He continues: Among some in EA and existential risk circles, my impression is that there is an unusual tendency to think that killing and violence can be morally right in various situations, and the people I have met and the statements I have seen in these circles appearing to be reasons for concern are more of a principled, dedicated, goal-oriented, chilling, analytical kind. Knutsson then remarks that “if I would do even more serious background research and start acting like some investigative journalist, that would perhaps increase the risk.” This stands out to me because I have done some investigative journalism. In addition to being noisy about the dangers of TESCREALism, I was the one who stumbled upon Bostrom’s email from 1996 in which he declared that “Blacks are more stupid than whites” and then used the N-word. This email, along with Bostrom’s “apology”—described by some as a flagrant “non-apology”—received attention from international media outlets. For utopians, critics aren’t mere annoyances, like flies buzzing around one’s head. They are profoundly immoral people who block the path to utopia, threatening to impede the march toward paradise, arguably the greatest moral crime one could commit. Even people within the TESCREAL community have become scared to speak out, for fear of retaliation. Last January, a group of about ten Effective Altruists (EAs) posted a lengthy critique of the movement anonymously, because of “the significant risk” that including their names “would pose to their careers, access to EA spaces, and likelihood of ever getting funded again.” When people inside a community become too afraid to publicly criticize it, that community starts to look rather like a cult. Indeed, last year I asked someone who was very prominent in the TESCREAL scene of San Francisco about the mood there, and they said that it had become a “full grown apocalypse cult.” The danger of secular apocalyptic cults is that, when members believe that the promises of utopia are about to be shattered, they may resort to extreme measures to keep those promises alive. The threats that I’ve received, the worries expressed by Knutsson, and the fact that TESCREALists themselves feel the need to hide their identities further bolsters my claim that this movement is dangerous. It operates like a cult, has “charismatic” leaders like Yudkowsky and Bostrom, and appears to be increasingly at ease with extreme rhetoric about how to stop the AGI apocalypse. The warnings I articulated in 2019 and 2021 are no longer merely hypothetical. What we’re seeing now is exactly what I worried could happen in the years to come. It didn’t take long for my predictions to prove accurate. When leading TESCREALists argue that existential risks are the “one kind of catastrophe that must be avoided at any cost,” to quote Bostrom with italics added, one shouldn’t be surprised if members of the community start talking about the use of force, military strikes, or targeted killings to reduce the supposed “existential risk” of AGI. I do not know how this ends, but what’s clear from history is that many utopian movements, embracing a kind of “utilitarian” reasoning, have left a trail of destruction behind them. Will this time be any different? It depends on whether the power and influence of the TESCREAL movement continues to grow, and right now, the trendlines are ominous.

#### AI doomers’ techno-paradise is contingent on capitalistic exploitation and elimination of those not deemed worthy of salvation – AI existential risk literature assumes a future for whites, not humanity

Torres 23 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 6-30-2023, "Does AGI Really Threaten the Survival of the Species?", Truthdig, https://www.truthdig.com/articles/does-agi-really-threaten-the-survival-of-the-species/]/Kankee

\*note: TESCREAL stands for transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism

Have you heard the news? Artificial general intelligence, or AGI, is going to kill us all. Even people living in the world’s most remote regions — the Amazon rainforest, northern Siberia and Antarctica. According to “AI doomers,” a “misaligned” AGI will annihilate the entire human population not because it hates us or has a ghoulish lust for omnicidal violence, but simply because human beings “are made out of atoms which it can use for something else.” We are resources that it will utilize to achieve its ultimate goals, whatever they happen to be. Or so the “doomers” claim. This idea of AGI posing an “existential risk” to humanity has gained a lot of steam following the release of ChatGPT last November. Geoffrey Hinton, the “godfather of AI,” has expressed this worry, as has the self-described “decision theorist” Eliezer Yudkowsky, who argued in Time magazine that we should be willing to risk thermonuclear war to prevent AGI in the near future. A more recent article from OpenAI, the for-profit company behind ChatGPT and GPT-4, cites “existential risks” in calling for an international agency to “regulate” AGI research. But just how plausible are these “existential risks” from AGI? What exactly is an “existential risk,” and why is it so important to avoid? The first thing to note is that “existential risk” has both colloquial and canonical definitions. In the colloquial sense, it simply refers to the destruction of our species, resulting in total human extinction, whereby Homo sapiens disappears entirely and forever. I suspect this is what most people think of when they hear the term “existential risk,” which makes sense given that “existential” means “of or relating to existence.” The canonical definition is far more controversial because it’s bound up with the “TESCREAL bundle” of ideologies that I discussed in my last article for this series. To recap briefly, the acronym “TESCREAL” stands for — and prepare yourself for a mouthful here — transhumanism, Extropianism, singularitarianism, cosmism, Rationalism, Effective Altruism and longtermism. The clunky acronym is important because these ideologies are hugely influential within Silicon Valley and the tech world more generally. Elon Musk is a TESCREAList, as is Sam Altman, the CEO of OpenAI, which Musk and Altman co-founded with money from Peter Thiel and others. Understanding this bundle is key to making sense of what’s going on with the tech elite, and it’s what gave rise to the canonical definition of “existential risk” in the early 2000s. On this definition, “existential risk” refers to any event that would prevent us from fulfilling “our longterm potential” in the universe. What does this mean? According to TESCREALists, fulfilling our “potential” would involve things like — and this is not an exaggeration — creating a new race of superior “posthumans,” colonizing the universe, building giant computer simulations full of trillions and trillions of “happy” digital people and ultimately generating “astronomical” amounts of “value” over the coming “millions, billions and trillions of years.” Put differently, the canonical definition of existential risk is anything that ruins our chances of realizing a techno-utopian world among the heavens populated by enormous numbers of digital posthumans awash in — as Nick Bostrom put it in his “Letter from Utopia” — “surpassing bliss and delight.” This is what many TESCREALists mean when they talk about “existential risks.” It’s why they argue that existential risks “must be avoided at any cost,” in Bostrom’s words, as failing to avoid them would mean the loss of utopia and astronomical value. It’s why Yudkowsky argued in his Time article that we should be willing to risk an all-out thermonuclear war to prevent apocalyptic AGI from being developed. The reasoning goes like this: thermonuclear war probably wouldn’t completely destroy our “vast and glorious” future in the universe, because there would almost certainly be some survivors who could rebuild society. According to one recent study, if a thermonuclear war were to break out between the U.S. and Russia tomorrow, a reassuring 3 billion people would remain — plenty enough to keep industrial civilization, our springboard to outer space, roaring. In contrast, doomers like Yudkowsky believe that an AGI apocalypse would kill every last person on Earth, thus forever erasing the techno-utopian world that we could have otherwise drawn. An AGI apocalypse would be an existential catastrophe, whereas a thermonuclear war almost certainly wouldn’t. Hence, when someone on Twitter posed the question, “How many people are allowed to die to prevent AGI?” Yudkowsky offered a jaw-dropping response: There should be enough survivors on Earth in close contact to form a viable reproductive population, with room to spare, and they should have a sustainable food supply. So long as that’s true, there’s still a chance of reaching the stars someday. This points to why I’ve argued that you shouldn’t care about “existential risks” in the canonical sense: the utopian vision at the heart of this concept is a total nonstarter, and could even be dangerous. Who cares if 10^58 digital people — Bostrom’s actual estimate — living in huge computer simulations spread throughout the universe exist or not? Many TESCREALists would say that the nonbirth of these people would constitute a horrendous moral catastrophe. But would it? I’d say no, and I think most people agree with me. If this “utopia” is a nonstarter, then who cares about “existential risks” on the canonical account? Yet the situation is more insidious than this, because both the pursuit and realization of this “utopia” would almost certainly have catastrophic consequences for most of humanity. Consider OpenAI. As noted above, the company was founded by TESCREALists and is now run by one. Its explicit mission is “to ensure that artificial general intelligence benefits all of humanity.” Yet the large language models (LLMs) behind ChatGPT and GPT-4 were built on massive amounts of intellectual property theft. This is why a number of lawsuits have recently been filed against OpenAI. Even more, in curating the training data for its LLMs, OpenAI hired a company that paid Kenyan workers as little as $1.32 per hour to sift through “examples of violence, hate speech and sexual abuse” on the Internet, which left some workers traumatized. Does this look like “benefitting all of humanity”? Obviously not, which illustrates how the pursuit of “utopia” will leave a trail of destruction in its wake: Some people, perhaps a very large number, are going to get trampled in the march to paradise. Now let’s imagine that TESCREALists succeed in creating their “utopia.” Who exactly is this utopia for? Will some people get left out? If so, which people? One of the most striking things about the TESCREAL literature is that there’s virtually zero consideration of what the future should look like from nonWestern perspectives. When TESCREALists at the influential Future of Humanity Institute, founded by Bostrom in 2005, talk about “the future of humanity,” they aren’t talking about the interests of “humanity” as a whole. Rather, they’re promoting an extremely narrow, thoroughly Western vision founded on the Baconian dream of plundering the universe and the capitalistic desideratum of maximizing economic productivity to the limits. In fact, Bostrom once literally defined an “existential risk” as anything that would prevent us from reaching “technological maturity,” defined as “the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved.” By subjugating nature and maximizing productivity, he reasons, we would be optimally positioned to realize the TESCREAL vision of a posthuman paradise. Absent is any reference whatsoever to alternative conceptions of the future from the perspective of, say, Muslims, Afrofuturism, feminism, queerness, disability and so on. Also missing is any discussion of which future might be best for the nonhuman creatures with whom we share planet Earth. Prominent TESCREAList William MacAskill even argues that our destruction of the natural world might very well be net positive. Why? Because wild animals suffer, and fewer wild animals means less wild-animal suffering. In his words, if we assess the lives of wild animals as being worse than nothing on average, which I think is plausible (though uncertain), then we arrive at the dizzying conclusion that from the perspective of the wild animals themselves, the enormous growth and expansion of Homo sapiens has been a good thing. Far from being “utopia,” the grandiose fantasies at the heart of TESCREALism look more like a dystopian nightmare, and for many groups of people it might even entail their elimination. Almost every imagined utopia leaves some people out — that’s the nature of utopia — and so far as I can tell, if the TESCREAL vision were fully realized, most of humanity would lose (to say nothing of the biosphere more generally). The canonical definition of “existential risks” is more than a nonstarter, then: it’s built around a deeply unappealing vision of the future designed by privileged white men at elite universities and within Silicon Valley, which they now want to impose on everyone else, whether we like it or not. But what of “existential risks” in the colloquial definition, simply meaning “the annihilation of every human on Earth”? Should you care if AGI poses an existential risk in this sense? The answer is “Of course you should, since human extinction would mean that you, your family and friends and everyone else would die a terrible death.” Does this mean that we really ought to take Yudkowsky and the other “doomer” TESCREALists seriously? The answer depends on how plausible an AGI apocalypse is. Not that long ago, I found the arguments for why a “misaligned” AGI would destroy us somewhat convincing, although I have recently changed my tune. A major problem with these arguments is that it’s not clear how exactly an AGI could actually kill everyone on Earth. In his 2014 bestseller “Superintelligence,” Bostrom outlines a situation in which AGI “strikes” humanity “through the activation of some advanced weapons system that the AI has perfected.” He continues: If the weapon uses self-replicating biotechnology or nanotechnology, the initial stockpile needed for global coverage could be microscopic: a single replicating entity would be enough to start the process. In order to ensure a sudden and uniform effect, the initial stock of the replicator might have been deployed or allowed to diffuse worldwide at an extremely low, undetectable concentration. At a pre-set time, nanofactories producing nerve gas or target-seeking mosquito-like robots might then burgeon forth simultaneously from every square meter of the globe. Yudkowsky points to a different set of possibilities. After being asked how an AGI could obliterate humanity so that there are “no survivors,” he speculates that it could synthesize a pathogenic germ that “is super-contagious but not lethal.” Consequently, no significant efforts are [put] into stopping this cold that sweeps around the world and doesn’t seem to really hurt anybody. And then, once 80% of the human species has been infected by colds like that, it turns out that it made a little change in your brain somewhere. And now if you play a certain tone at a certain pitch, you’ll become very suggestible. So, virus-aided, artificial pathogen-aided mind control. Realizing that this wouldn’t necessarily kill everyone, Yudkowsky then proposed a different scenario. Imagine, he says, that an AGI hell-bent on destroying us creates, something that [can] reproduce itself in the air, in the atmosphere and out of sunlight and just the kind of atoms that are lying around in the atmosphere. Because when you’re operating at that scale, the world is full of an infinite supply of … perfect spare parts. Somebody calculated how long it would take “aerovores” to replicate and blot out the sun, use up all the solar energy. I think it was like a period of a couple of days. At the end of all this is tiny diamondoid bacteria [that] replicate in the atmosphere, hide out in your bloodstream. At a certain clock tick, everybody on Earth falls over dead in the same moment. Does any of this make sense, or does it sound like utter madness? As Carl Sagan famously declared, extraordinary claims require extraordinary evidence, and the doomer’s claim that a “misaligned” AGI will kill the entire human population is about as extraordinary as they get. So where is the evidence? The scenarios above are hardly convincing. The doomers have two responses to this. The first is based on the third of Arthur C. Clarke’s “three laws,” which states: “Any sufficiently advanced technology is indistinguishable from magic.” We can illustrate this by analogy: picture yourself traveling back in time, convincing a Neanderthal to hop in the time machine with you, and then zipping back to the present. Imagine what this Neanderthal would think about our world. Airplanes, brain surgery, smartphones and GPS would all appear to be “magic.” AI doomers could use this to claim that: “Look, maybe we can’t come up with a plausible way that AGI would kill everyone. But that’s no reason to reject our doomsday hypothesis. By definition, a superintelligent AGI would be unimaginably ‘smarter’ than us. This would enable it to create super-advanced technologies, and super-advanced technologies would be able to manipulate and rearrange the world in ways that are, from our limited human point of view, indistinguishable from magic.” But if we’re talking about a “god-like AI” that traffics in “scientific magic,” then literally anything goes. There are no longer any rules to the conversational game, no sense to make of this post-AGI world. Maybe the AGI magically destroys us — but since we’re in the land of magic, maybe it doesn’t. The doomers would point out that lots of potential scenarios all lead to the conclusion that the AGI would try to kill us, but the arguments for this claim are all formulated within mere human frameworks. Perhaps these frameworks are wrong in ways that we simply cannot understand, just as a dog will never make sense of its owner’s behavior by extrapolating its own thoughts and instincts into the realm of human action. If we’re talking about magic beyond our comprehension, then there’s just nothing meaningful to say, and hence these doomsday worries aren’t based on anything intelligible. The second response to the “extraordinary evidence” objection involves “expected value.” AI doomers might say that even if these apocalyptic scenarios are extremely unlikely, the losses that our extinction would entail are so great that we should still take them very seriously. In other words, if a risk is the probability of an outcome multiplied by its badness, and if the badness of an outcome is enormous, then the risk could be very large even if the probability is miniscule. Maybe “target-seeking mosquito-like robots” and atmospheric self-replicating “diamondoid bacteria” are highly improbable, but the stakes are so massive — everyone on Earth dying — that we should still worry about the AGI apocalypse. The problem with this is that the very same argument could be used to support all manner of absurd conclusions. What if I told you that the next time you sneeze, the universe will explode in a giant, violent conflagration? You might think that this warning is ridiculous, lacking any explanation of a causal mechanism. But if you take seriously the “expected value” definition of “risk” mentioned above, then you might have a pretty good “reason” never to sneeze again. In fact, TESCREALists are aware of this problem: they call it “Pascal’s mugging,” which describes a situation in which a mugger says that if you don’t give him $5, he’ll torture trillions of people in a parallel universe. Although his claim is almost certainly nonsense, you can’t know for sure that he’s lying — the universe is, after all, a very strange place! So, you crunch the numbers, and ultimately hand over your $5. The mugger used expected value theory as his weapon, and it worked. Where does all this leave us? On the one hand, you shouldn’t care about “existential risks” in the canonical sense used by TESCREALists. This is because the concept is intimately bound up with a deeply problematic vision of utopia built on Western, Baconian-capitalistic values of exploitation, subjugation and maximization. Not only is this vision of 10^58 people in huge computer simulations a complete nonstarter, but its pursuit and realization would be catastrophic for most of humanity. The utopian world of TESCREALism would be a dystopia for everyone else. On the other hand, you should care about the possibility of an AGI causing an “existential” catastrophe if understood colloquially as total human extinction. The problem, though, is that there’s no plausible account of how an AGI could realistically accomplish this, and claiming that it would employ “magic” that we just can’t understand essentially renders the whole conversation vacuous, since once we’ve entered the world of magic, anything goes. To repurpose a famous line from Ludwig Wittgenstein: “What we cannot speak about we must pass over in silence.” This is why I’ve become very critical of the whole “AGI existential risk” debate, and why I find it unfortunate that computer scientists like Geoffrey Hinton and Yoshua Bengio have jumped on the “AI doomer” bandwagon. We should be very skeptical of the public conversation surrounding AGI “existential risks.” Even more, we should be critical of how these warnings have been picked up and propagated by the news, as they distract from the very real harms that AI companies are causing right now, especially to marginalized communities. If anything poses a direct and immediate threat to humanity, it’s the TESCREAL bundle of ideologies that’s driving the race to build AGI, while simultaneously inspiring the backlash of AI doomers who, like Yudkowsky, claim that AGI must be stopped at all costs — even at the risk of triggering a thermonuclear war.

#### AI risk mitigation undermines poverty reduction, treating the ultra-poor as disposable and dis-valuable compared to future beings

Torres 23 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 3-30-2023, "Giving Altruism a Bad Name", Truthdig, https://www.truthdig.com/articles/giving-altruism-a-bad-name/]/Kankee

The effective altruist (EA) movement is young. It was only around 2010 that it started to take shape, yet over the past 13 years it’s become a “powerful global network of think tanks, nonprofit organizations and wealthy donors that dole out hundreds of millions of dollars in annual charitable donations.” If you were to ask someone from the EA community, they’d say this is great news. In their telling of the story, the movement “has saved hundreds of thousands of lives” and “made the world a lot better.” But is this true? In a recent book titled “The Good It Promises, the Harm It Does,” an impressive group of scholars and activists make the case that EA has, in fact, left quite a bit of wreckage in its wake. One chapter notes that “#MeToo advocates [have] described how the funding metrics of EA ignored hostile work environments, and how many have been hurt by known serial sexual exploiters who lead groups assessed as ‘effective’ by EA-tied groups.” A recent Time magazine article explores precisely this topic, reporting on several women who describe, in disturbing detail, how the EA community “has a toxic culture of sexual harassment and abuse.” The problem is, apparently, pervasive. Another author, Brenda Sanders, reports that she “was once told that a prominent animal rights movement donor would never fund my work because ‘there’s no way to prove how effective it is.’” Her work focuses on vegan activism in low-income communities of color, which involves “hosting workshops, classes, festivals, food tastings, film screenings and other events that introduce people in low-income communities of color to the benefits of making healthier, kinder, more sustainable choices.” After rejecting what she calls the “white-centric view of activism,” i.e., the idea that one should strive to maximize one’s “return on investment,” Sanders has a few choice words for this unnamed donor, “some rich white guy.” “I would also point out,” she writes, “that refusing to support work being done by a Black activist in Black communities is upholding white supremacist ideas about which communities are worthy of support and which ones aren’t. In other words, it’s racist, plain and simple.” In an earlier article I wrote for Truthdig, racism was a prominent theme. Pretty much everywhere one peeks around the neighborhood of EA, one discovers hints of racism — to say nothing of ableism, classism and sexism. After revelations last month that Nick Bostrom, one of the most influential philosophers within EA, wrote in old email claiming that “Blacks are more stupid than whites,” followed by the N-word, it became clear that many EAs are actually sympathetic with the view that racial differences in “intelligence” do exist, or at least aren’t particularly bothered by his claim. Others in the community have approvingly cited the work of Charles Murray, world-famous for his scientific racism, and one of the main focuses of EA — so-called “longtermism” — traces its roots right back to the 20th-century eugenics movement. None of this would be particularly noteworthy (the world is full of rotten ideas) except for the fact that EA and its longtermist offshoot have become extremely influential over the past decade. In this piece for my Truthdig series “Eugenics in the Twenty-First Century: New Names, Old Ideas,” I’d like to take a closer look at how EA actually causes more harm than good, and how it’s approach to philanthropy further entrenches the political system of white supremacy — i.e., the continuing dominance of white people within the power structures that define our world today. Our case study will be that of Anthony Kalulu, an Ugandan farmer who has, in his words, “spent the vast portion of my life in ultra poverty.” He’s never “stepped a foot outside of Sub Saharan Africa, the world’s epicenter for chronic poverty,” and even as recently as 2015 he “was still going entire days without food”— a situation he’s been dealing with “since my years in childhood.” I first became aware of Kalulu last December, after the renowned AI researcher Timnit Gebru tweeted about some of his work. Since then, Kalulu and I have become friends, and I have donated twice to his grassroots organization the Uganda Community Farm (UFC), which we’ll get to below. In a series of poignant and insightful articles on his website — as well as three for The Guardian — Kalulu makes the case that traditional philanthropy simply doesn’t work. As I understand Kalulu’s argument, what happens is this: charities from the Global North adopt a “top-down approach” to alleviating poverty that doesn’t take seriously the unique challenges, struggles and circumstances of particular regions of the world. With a kind of “one size fits all” mentality, they tend to impose programs or strategies mostly devised by white people in the Global North, and fail to properly recognize that the root causes of poverty may differ from one region to the next. Without addressing these unique-to-the-region causes, living conditions might improve a bit while the charity is present, but once they’re gone, everything falls apart again. Fundamental, lasting change never happens. Kalulu thus argues that this whole paradigm needs to shift. What’s needed is a “bottom-up approach” that takes seriously the knowledge that local people have accumulated of the specific drivers of poverty in their particular regions. In practice, this would mean directly funding grassroots organizations run by locals — people with intimate, first-hand, ground-level knowledge of what’s going on. It’s precisely this knowledge that would enable such organizations to develop anti-poverty programs that would be custom designed to address the root causes of a community’s problem, thus enabling that community to pull itself out of poverty once and for all. As Kalulu puts it, traditional philanthropy pushes “their own predetermined solutions, which aren’t rooted in the lived experiences of the extreme poor,” and hence aren’t able to fix, in any permanent way, what’s broken. Not every region is the same, a fact that traditional philanthropic approaches neglect with their “one- size-fits-all” approach. Another way to think about this goes as follows: due in part to the ongoing legacies and influence of colonialism, exploitation and extractive capitalism, the Global North has most of the power and wealth. By utilizing a top-down approach, people in the Global North get to retain this power. That is to say, they maintain power over who gets wealth transferred to them and how this wealth is transferred. The decision-making process remains in the hands of mostly white people in the Global North, who live literally thousands of miles away from countries like Uganda and have no first-hand knowledge of what, say, Ugandan farmers are actually struggling with. Consequently, the traditional approach is profoundly disempowering to those it aims to help: recipients of charity become merely passive participants in trying to solve their problems, rather than decision-makers in their own right. Perhaps you can see how this further entrenches the political system of white supremacy. It also explains why philanthropic efforts of the past have so often failed to extricate people from the cycle of poverty: they aren’t flexible enough, and don’t have enough knowledge of the unique circumstances of specific communities, to implement programs that can address the root causes of destitution. This is why Kalulu argues that we need a new paradigm: a bottom-up approach that transfers not only wealth but power to those suffering from extreme poverty, thus empowering these people to finally fix what’s broken in their communities, regions, local economies, political systems and so on. Should money be spent building a new school, highway or food processing plant? Should cash be transferred straight into the pockets of local individuals, as the charity GiveDirectly does? Is the best use of funds buying anti-malaria bed nets? Perhaps those actually living these realities are best positioned to answer such questions. Right now, though, these people have little or no voice. They have no real say in matters that directly affect them, their communities and their collective futures. EA enters the picture because, in Kalulu’s words, it’s “even worse than traditional philanthropy in the way it excludes those of us who are directly battling ultra poverty in the Global South.” There are two interrelated reasons for this: first, EA makes it even less likely that grassroots organizations will ever get funded. The reason is that the EA movement insists that members should donate exclusively to those causes, charities or organizations deemed “effective” according to certain metrics or criteria, developed by EA organizations like GiveWell, which is based in San Francisco. If an organization doesn’t qualify as “effective,” then it doesn’t deserve funding. This is what “doing good better” means: funneling money to those, and only those, charities that get the biggest bang for one’s buck. As Kalulu makes this point, in the name of being “effective,” EA has … indoctrinated its followers to strictly support a small, select list of charities that have been labeled “most effective” by the movement’s own charity raters like GiveWell, Giving What We Can, The Life You Can Save, etc., of which the named charities, right now, are all Western. … [I]n the eyes of a true effective altruist, lending a hand directly to people like us, is not one of those ways of doing the most good, unless we are part of a charity that has been declared by organizations like GiveWell as “effective” — which, at least for now, means being part of a Western charity. Consequently, people like Kalulu, who are pleading for money to support their grassroots organizations, are left with even less hope than before. At least with traditional philanthropy, some grassroots organizations occasionally will get funded. “For traditional philanthropy, albeit being a sector (or a community) that almost never supports the poor directly,” Kalulu writes, “there are countless occasions where even organizations like the Gates Foundation have funded the smallest grassroots organizations in the Global South. Not the case with effective altruism.” The second problem concerns the metrics that EA uses, which are designed such that many grassroots organizations in the Global South could probably never make the cut. As Alice Crary observes in her chapter of “The Good It Promises, the Harm It Does,” “EA’s metrics are best suited to detect the short-term impact of particular actions, so its tendency to discount the impact of coordinated actions can be seen as reflecting ‘measurability bias.’” Although her focus here is social movements, in particular, the same idea applies to the case at hand: grassroots efforts to implement programs that produce long-lasting, rather than merely short-term or immediate, improvements in specific communities or regions. For example, GiveWell calculates that just $3,340 can save the life of a child if donated to the Against Malaria Foundation, which distributes bed nets to people in regions of the world where malaria is prevalent. This is an easy-to-measure return on one’s investment, alluding once more to the “white-centric view of activism” that Sanders mentions above. Now consider Kalulu’s own grassroots organization, the Uganda Community Farm (UCF). One of its central goals is to build an agro-processing plant that would enable farmers in the area to establish reliable links to ready markets. One reason farmers in the area are poor, Kalulu reports, is that while they have produce to sell, they don’t have any consistent way to actually sell it. An agro-processing plant would provide the infrastructure necessary for farmers to climb out, and stay out, of poverty. But notice how difficult it would be to measure the benefits of this plant, especially in the short term. Exactly how many lives would be saved? Would the agro-processing plant save more children per $3,340 than the Against Malaria Foundation? The “return” for investing in UCF’s project isn’t easily measured by EA’s metrics, and this puts it at a profound disadvantage; indeed, it’s not clear that it could ever qualify as “effective” compared to, say, distributing bed nets, given the “measurability bias” of EA’s framework. But perhaps this isn’t a problem with grassroots organizations. It’s a problem with the metrics used to determine whether a cause is worthy of being funded. “As someone who lives in one of the poorest corners of the planet,” Kalulu writes, I can tell you that EA’s idea of “effective” is mostly correct only from a Western viewpoint, and again, it is all because the global development community as a whole, EA included, has chosen to make the extreme poor passive participants in creating the change they want to see. If you visited a truly impoverished country like Uganda, you will quickly notice that many of the things that effective altruists call “effective” … are the same short-term, disposable solutions that have not only kept their recipients in abject poverty, but also, they are the very kind of solutions that often disappear the same day their proponents exit. These major shortcomings of the EA approach to philanthropy got personal for Kalulu in 2021 after he “contacted about a hundred people (mostly individuals) who identify as effective altruists, asking them not for money, but for social media posts about the plight of poverty in my region, and my quest to do something about it.” Posting something on social media costs nothing, yet only a single EA agreed to share a link to Kalulu’s Go Get Funding webpage. One EA responded to Kalulu’s request: “I usually share projects that have been vetted by GiveWell or a similar organization…Why do you think your project is more effective compared to all the others?”

So, it’s pretty clear that top-down philanthropy isn’t working, and yet the EA movement — even more than traditional efforts — has embraced precisely this approach. Making matters worse, Kalulu gestures at the fact that a growing number of EAs are embracing longtermism, which, in its radical form, holds that we should focus our “charitable” efforts almost entirely on shaping the very far future, rather than on helping those in need today. The claim isn’t that people who might (or might not) exist in the future — thousands, millions, billions and trillions of years from now — are more important. The reasoning is that every person, whether they are an actual person or a merely possible person in the future, should count for one. And since there could be way more people in the future if we colonize the universe and create huge computer simulations full of digital beings, trying to positively influence their lives should take priority over the lives of people suffering right now. This is basically a numbers game. For example, imagine that there’s a 0.0000000000000001 chance that some action you take today could add one unit of happiness to 1,000,000,000,000,000,000 people who exist in the far future. This yields an “expected value,” as longtermists would say, of 100, by simply multiplying these two numbers together. Now imagine that there’s a different action you could take that would, with a probability of 0.5, add 100 units of happiness to one person right now — say, someone struggling to buy three meals a day. This yields an expected value of only 50. Since 100 is twice as big as 50, if you were forced to choose between these two actions, you should take the first one — not the second! This is how EA longtermists reason, and it’s what leads them to deprioritize the problem of global poverty, that is, relative to the importance of “helping” people in the far future — even if the probability of succeeding is extremely small. In fact, longtermists argue that one way to “help” these people is to ensure that they come to exist in the first place. On their view, which is highly contentious among philosophers, “could exist” implies “should exist” assuming that such people would have lives that are better than miserable. It follows that the non-existence of a potentially vast number of future people — many living in computer simulations — would constitute an enormous moral catastrophe, much greater than the catastrophe of global poverty in the present. Again, this is not because current people matter less, but because there are so many “happy” people who could (and therefore should) exist in the future. Since one way these hordes of unborn digital beings could fail to exist is if humanity were to go extinct, this line of reasoning leads longtermists to prioritize mitigating risks to our extinction over most everything else, even if these risks are highly speculative and very improbable. Hence, as Kalulu observes, longtermists “consider things like artificial intelligence as being existential threats to humanity (before they pose any threat), yet the movement won’t even lend a hand directly to those of us who are already starving in the present day.” This might sound outrageous, but it’s a straightforward implication of the longtermist worldview. To quote one of the leading longtermists, Hilary Greaves, on this point: There’s a clear case for transferring resources from the affluent Western world to the global poor. But longtermist lines of thought suggest that something else would be better still. There are a lot of candidates for potentially very high value longtermist interventions. … The most clear-cut one, I think, is reducing risks of premature human extinction. … Even if we can do anything that reduces the probability of premature human extinction by a tiny amount, in expected value terms, that is, when you average across your uncertainty, the contribution of that kind of intervention could be massive — much greater, even, than the best things we can do in the area of global poverty. These are some reasons that EA is even worse than traditional philanthropy. On the one hand, it hopelessly sidelines grassroots projects like Kalulu’s own UCF, and in doing so keeps all the decision-making power in the hands of mostly white people in the Global North. As Kalulu writes, “by deciding not to directly support the people who live in ultra poverty,” EA “not only squanders the chance to catalyze lasting, self-sustainable change, but also effectively puts the lived experiences of the world’s ultra poor to waste.” On the other hand, the EA movement has convinced an increasing fraction of its community over the past five years that global poverty as a whole isn’t the most “effective” cause area. Instead, we should redirect our finite resources to ensuring that humanity survives long enough to colonize space, spread throughout the accessible universe, populate exoplanets and build giant computer simulations in which trillions and trillions of digital people live supposedly “happy” lives. Although not all EAs have pivoted toward longtermism, and hence away from global poverty, the most prominent members of the EA community like Toby Ord and William MacAskill — both of whom once focused primarily on global poverty rather than the far future — have been vigorously promoting the longtermist perspective, and as I’ve discussed on many occasions before, longtermism is becoming hugely influential in the world more generally. The result is a truly heartbreaking situation for people like Kalulu. As he reports in his critique of EA, referencing the exchange with an EA mentioned earlier: Looking back, I couldn’t help but wonder how people like us will really ever escape poverty. I was talking with him right here in my village of Namisita (in Kagumba Sub County, Kamuli), a place where some households find it very hard to even afford soap, and a place where there is nothing whatsoever that is happening to end poverty.

#### The EA existential risk do-gooder complex brushes issues like sexual harassment under the rug – its deemed unimportant compared to “saving humanity”

Alter 23 [Charlotte Alter, senior correspondent at Time, 02-03-2023, “Effective Altruism Promises to Do Good Better. These Women Say It Has a Toxic Culture Of Sexual Harassment and Abuse,” Time, https://time.com/6252617/effective-altruism-sexual-harassment/]/Kankee

Keerthana Gopalakrishnan once considered herself an effective altruist. As a college student in India, she immersed herself in the social movement, reading its canonical texts like Doing Good Better, listening to its podcasts, and devouring effective altruism (EA) blogs in an attempt to figure out how to create a life of maximum moral impact. When the world started opening up from the COVID-19 pandemic, she moved to San Francisco and went to EA meetups, made friends with other EAs, and volunteered at EA conferences where they talked about how to use evidence and reason to do the most good in the world. But as Gopalakrishnan got further into the movement, she realized that “the advertised reality of EA is very different from the actual reality of EA,” she says. She noticed that EA members in the Bay Area seemed to work together, live together, and sleep together, often in polyamorous sexual relationships with complex professional dynamics. Three times in one year, she says, men at informal EA gatherings tried to convince her to join these so-called “polycules.” When Gopalakrishnan said she wasn’t interested, she recalls, they would “shame” her or try to pressure her, casting monogamy as a lifestyle governed by jealousy, and polyamory as a more enlightened and rational approach. After a particularly troubling incident of sexual harassment, Gopalakrishnan wrote a post on an online forum for EAs in Nov. 2022. While she declined to publicly describe details of the incident, she argued that EA’s culture was hostile toward women. “It puts your safety at risk,” she wrote, adding that most of the access to funding and opportunities within the movement was controlled by men. Gopalakrishnan was alarmed at some of the responses. One commenter wrote that her post was “bigoted” against polyamorous people. Another said it would “pollute the epistemic environment,” and argued it was “net-negative for solving the problem.” Gopalakrishnan is one of seven women connected to effective altruism who tell TIME they experienced misconduct ranging from harassment and coercion to sexual assault within the community. The women allege EA itself is partly to blame. They say that effective altruism’s overwhelming maleness, its professional incestuousness, its subculture of polyamory and its overlap with tech-bro dominated “rationalist” groups have combined to create an environment in which sexual misconduct can be tolerated, excused, or rationalized away. Several described EA as having a “cult-like” dynamic. Julia Wise, the longest-serving employee of the Centre for Effective Altruism (CEA), an Oxford, England-based charity responsible for growing and maintaining the EA community, acknowledges that there have been reports of sexual harassment within the community. But she questions whether the movement itself is responsible. Sexual misconduct is a problem throughout society, after all, and EA leaders cannot control the behavior of everyone moving in and around it. “Some of the concerns that have come up are maybe made by people in EA, but the perpetrator attended an event a couple years ago but they’re not that involved,” says Wise. “How do you figure out what is a community problem versus what is a Bay Area problem or sex problem or something else?” This story is based on interviews with more than 30 current and former effective altruists and people who live among them. Many of the women spoke on condition of anonymity to avoid personal or professional reprisals, citing the small number of people and organizations within EA that control plum jobs and opportunities. Much of the alleged abuse they detailed was concentrated in the Bay Area, but the women also described incidents that took place in three other states as well as overseas. Many of them asked that their alleged abusers not be named and that TIME shield their identities to avoid retaliation. Their accounts were corroborated by other parties to the incidents, by people to whom the women spoke shortly afterward, and by contemporaneous documents and screenshots. While a few women have raised these issues on online forums, many spoke to TIME about their experiences with sexual misconduct in EA communities for the first time. One recalled being “groomed” by a powerful man nearly twice her age who argued that “pedophilic relationships” were both perfectly natural and highly educational. Another told TIME a much older EA recruited her to join his polyamorous relationship while she was still in college. A third described an unsettling experience with an influential figure in EA whose role included picking out promising students and funneling them towards highly coveted jobs. After that leader arranged for her to be flown to the U.K. for a job interview, she recalls being surprised to discover that she was expected to stay in his home, not a hotel. When she arrived, she says, “he told me he needed to masturbate before seeing me.” Several women say that the way their allegations were received by the broader EA community was as upsetting as the original misconduct itself. “The playbook of these EAs is to discourage victims to seek any form of objective, third-party justice possible,” says Rochelle Shen, who ran an EA-adjacent event space in the Bay Area and says she has firsthand experience of the ways the movement dismisses allegations. “They want to keep it all in the family.” In recent years, effective altruism morphed from a niche philanthropic community devoted to addressing worldwide poverty into a powerful global network of think tanks, nonprofit organizations and wealthy donors that dole out hundreds of millions of dollars in annual charitable donations. The movement has grown rapidly, with monthly active users on the EA forum growing fivefold since 2019, more than 6,000 attendees at EA global conferences in 2022, and at least 371 active EA chapters across more than 40 countries. Most of the movement’s members—who are overwhelmingly white, more than 70% male, and skew young, according to a recent survey of members of the community in 2020—are idealists drawn to the promise of building a better world by applying rigorous logic to moral decisions. Thousands have signed a pledge to tithe at least 10% of their income to high-impact charities. From college campuses to Silicon Valley startups, adherents are drawn to the moral clarity of a philosophy dedicated to using data and reason to shape a better future for humanity. Effective altruism has become something of a secular religion for the young and elite. But the fall of Sam Bankman-Fried, EA’s billionaire patron and most famous acolyte, who is now facing federal fraud charges tied to the collapse of his cryptocurrency exchange FTX, has put effective altruism under increased scrutiny. Like other recent social movements spanning the political spectrum, EA is diffuse and deliberately amorphous; anybody who wants to can call themselves an EA. And even in a community of self-styled do-gooders, “there certainly have been cases where people were treated badly, including sexual harassment,” says Wise, of the Centre for Effective Altruism. “This is an essential problem that all social groups face.” Wise, whose role at CEA involves overseeing community well-being, tells TIME she has fielded roughly 20 complaints per year in her seven years on the job, ranging from uncomfortable comments to more serious allegations of harassment and more. But with no official leadership structure, no roster of who is and isn’t in the movement, and no formal process for dealing with complaints, Wise argues, it’s hard to gauge how common such issues are within EA compared to broader society. The women who spoke to TIME counter that the problem is particularly acute in EA. The movement’s high-minded goals can create a moral shield, they say, allowing members to present themselves as altruists committed to saving humanity regardless of how they treat the people around them. “It’s this white knight savior complex,” says Sonia Joseph, a former EA who has since moved away from the movement partially because of its treatment of women. “Like: we are better than others because we are more rational or more reasonable or more thoughtful.” The movement “has a veneer of very logical, rigorous do-gooderism,” she continues. “But it’s misogyny encoded into math.” In a fashionable neighborhood of San Francisco, there is a Victorian house where a group of self-identified EAs, AI researchers, rationalist tech bros, and young women founders all sheltered together from the blistering Bay Area rents. While it had no formal relationship with effective altruism, roughly a third of the residents were EAs, and the house regularly hosted EA events. The residence was run by two co-leaders, a man and a woman, who signed the lease, managed the rent money, and handled the logistics of moving people in and out of a community that resembled a tech-era version of the 1960s Bay Area communes.

#### AI existential risk saviorism causes mass sexual abuse – risk researchers care not for women in light of saving the world

**Huet 23** [Ellen Huet, journalist for Bloomberg News and Bloomberg Businessweek and BA in English and Political Science at Stanford University, 3-6-2023, "The Real-Life Consequences of Silicon Valley’s AI Obsession", Bloomberg, https://www.bloomberg.com/news/features/2023-03-07/effective-altruism-s-problems-go-beyond-sam-bankman-fried]/Kankee

Taylor’s experience wasn’t an isolated incident. It encapsulates the cultural motifs of some rationalists, who often gathered around MIRI or CFAR employees, lived together, and obsessively pushed the edges of social norms, truth and even conscious thought. They referred to outsiders as normies and NPCs, or non-player characters, as in the tertiary townsfolk in a video game who have only a couple things to say and don’t feature in the plot. At house parties, they spent time “debugging” each other, engaging in a confrontational style of interrogation that would supposedly yield more rational thoughts. Sometimes, to probe further, they experimented with psychedelics and tried “jailbreaking” their minds, to crack open their consciousness and make them more influential, or “agentic.” Several people in Taylor’s sphere had similar psychotic episodes. One died by suicide in 2018 and another in 2021. Several current and former members of the community say its dynamics can be “cult-like.” Some insiders call this level of AI-apocalypse zealotry a secular religion; one former rationalist calls it a church for atheists. It offers a higher moral purpose people can devote their lives to, and a fire-and-brimstone higher power that’s big on rapture. Within the group, there was an unspoken sense of being the chosen people smart enough to see the truth and save the world, of being “cosmically significant,” says Qiaochu Yuan, a former rationalist. Yuan started hanging out with the rationalists in 2013 as a math Ph.D. candidate at the University of California at Berkeley. Once he started sincerely entertaining the idea that AI could wipe out humanity in 20 years, he dropped out of school, abandoned the idea of retirement planning, and drifted away from old friends who weren’t dedicating their every waking moment to averting global annihilation. “You can really manipulate people into doing all sorts of crazy stuff if you can convince them that this is how you can help prevent the end of the world,” he says. “Once you get into that frame, it really distorts your ability to care about anything else.” That inability to care was most apparent when it came to the alleged **mistreatment** of women in the community, as opportunists used the prospect of impending doom to **excuse vile** acts of **abuse**. Within the subculture of rationalists, EAs and AI safety researchers, sexual harassment and abuse are **distressingly** common, according to interviews with eight women at all levels of the community. Many young, ambitious women described a similar trajectory: They were initially drawn in by the ideas, then became immersed in the social scene. Often that meant attending parties at EA or rationalist group houses or getting added to jargon-filled Facebook Messenger chat groups with hundreds of like-minded people. The eight women say casual misogyny threaded through the scene. On the low end, Bryk, the rationalist-adjacent writer, says a prominent rationalist once told her condescendingly that she was a “5-year-old in a hot 20-year-old’s body.” Relationships with much older men were common, as was polyamory. Neither is inherently harmful, but several women say those norms became tools to help influential older men get more partners. Keerthana Gopalakrishnan, an AI researcher at Google Brain in her late 20s, attended EA meetups where she was hit on by partnered men who lectured her on how monogamy was outdated and nonmonogamy more evolved. “If you’re a reasonably attractive woman entering an EA community, you get a ton of sexual requests to join polycules, often from poly and partnered men” who are sometimes in positions of influence or are directly funding the movement, she wrote on an EA forum about her experiences. Her post was strongly downvoted, and she eventually removed it. The community’s guiding precepts could be used to justify this kind of behavior. Many within it argued that rationality led to superior conclusions about the world and rendered the moral codes of NPCs obsolete. Sonia Joseph, the woman who moved to the Bay Area to pursue a career in AI, was encouraged when she was 22 to have dinner with a 40ish startup founder in the rationalist sphere, because he had a close connection to Peter Thiel. At dinner the man bragged that Yudkowsky had modeled a core HPMOR professor on him. Joseph says he also argued that it was normal for a 12-year-old girl to have sexual relationships with adult men and that such relationships were a noble way of transferring knowledge to a younger generation. Then, she says, he followed her home and insisted on staying over. She says he slept on the floor of her living room and that she felt unsafe until he left in the morning. On the extreme end, five women, some of whom spoke on condition of anonymity because they fear retribution, say men in the community committed sexual assault or misconduct against them. In the aftermath, they say, they often had to deal with professional repercussions along with the emotional and social ones. The social scene overlapped heavily with the AI industry in the Bay Area, including founders, executives, investors and researchers. Women who reported sexual abuse, either to the police or community mediators, say they were branded as trouble and **ostracized** while the men were **protected**. In 2018 two people accused Brent Dill, a rationalist who volunteered and worked for CFAR, of abusing them while they were in relationships with him. They were both 19, and he was about twice their age. Both partners said he used drugs and emotional manipulation to pressure them into extreme BDSM scenarios that went far beyond their comfort level. In response to the allegations, a CFAR committee circulated a summary of an investigation it conducted into earlier claims against Dill, which largely exculpated him. “He is aligned with CFAR’s goals and strategy and should be seen as an ally,” the committee wrote, calling him “an important community hub and driver” who “embodies a rare kind of agency and a sense of heroic responsibility.” (After an outcry, CFAR apologized for its “terribly inadequate” response, disbanded the committee and banned Dill from its events. Dill didn’t respond to requests for comment.) Rochelle Shen, a startup founder who used to run a rationalist-adjacent group house, heard the same justification from a woman in the community who mediated a sexual misconduct allegation. The mediator repeatedly told Shen to keep the possible repercussions for the man in mind. “You don’t want to ruin his career,” Shen recalls her saying. “You want to think about the consequences for the community.” One woman in the community, who asked not to be identified for fear of reprisals, says she was sexually abused by a prominent AI researcher. After she confronted him, she says, she had job offers rescinded and conference speaking gigs canceled and was disinvited from AI events. She says others in the community told her allegations of misconduct harmed the advancement of AI safety, and one person suggested an agentic option would be to kill herself. For some of the women who allege abuse within the community, the most devastating part is the disillusionment. Angela Pang, a 28-year-old who got to know rationalists through posts on Quora, remembers the joy she felt when she discovered a community that thought about the world the same way she did. She’d been experimenting with a vegan diet to reduce animal suffering, and she quickly connected with effective altruism’s ideas about optimization. She says she was assaulted by someone in the community who at first acknowledged having done wrong but later denied it. That backpedaling left her feeling doubly violated. “Everyone believed me, but them believing it wasn’t enough,” she says. “You need people who care a lot about abuse.” Pang grew up in a violent household; she says she once witnessed an incident of domestic violence involving her family in the grocery store. Onlookers stared but continued their shopping. This, she says, felt much the same. None of the abuse alleged by women in the community makes the idea of AI safety less important. We already know all the ways that today’s single-tasking AI can distort outcomes, from racist parole algorithms to sexist pay disparities. Superintelligent AI, too, is bound to reflect the biases of its creators, for better and worse. But the possibility of marginally safer AI doesn’t make women’s safety less important, either. Twenty years ago, Yudkowsky’s concerns about AI safety were fringe. Today, they have billions of dollars behind them and more piling up—Google invested $400 million in Anthropic in February—but safety-focused efforts remain a tiny sliver of the money the industry is dedicating to the evolving AI arms race. OpenAI’s ChatGPT can pass a law school exam; its DALL-E can paint you pink dolphins leaping through clouds. Microsoft is piloting AI in Bing search. Even though the consensus view is that truly sentient AI remains a ways off, the pace of research and advancement is rapidly accelerating. The questions that haunt the movement are becoming more relevant, as are its sins. Yudkowsky now views OpenAI’s commercial efforts as “nearly the worst possible” path, one that will hasten our doom. On a podcast last month, he said he’d lost almost all hope that the human race could be saved. “The problem is that demon summoning is easy, and angel summoning is much harder,” he said. In 2003, around the time the Matrix sequels were in theaters and AI doomsday scenarios were mostly relegated to late-night dorm talk, Bostrom proposed a thought experiment about an AI whose only goal is to make the largest possible number of paper clips. That AI would quickly realize that to maximize its goal, there should be no humans: If humans decided to switch off the AI, that would prevent paper clip creation, and humans contain atoms that could be made into paper clips. The AI, he concluded, would be strongly incentivized to find ways to strip-mine us and everything else on the planet to reach its goal. The paper clip maximizer, as it’s called, is a potent meme about the pitfalls of maniacal fixation. Every AI safety researcher knows about the paper clip maximizer. Few seem to grasp the ways this subculture is mimicking that tunnel vision. As AI becomes more powerful, the stakes will only feel higher to those **obsessed** with their self-assigned quest to keep [AI] under rein. The collateral damage that’s already occurred **won’t** matter. They’ll be thinking only of their own kind of paper clip: saving the world.

#### Effective altruism leads to misogyny and justification of human rights erosion.

**Huet 23** [Ellen Huet, journalist for Bloomberg News and Bloomberg Businessweek and BA in English and Political Science at Stanford University, 3-6-2023, "The Real-Life Consequences of Silicon Valley’s AI Obsession", Bloomberg, https://www.bloomberg.com/news/features/2023-03-07/effective-altruism-s-problems-go-beyond-sam-bankman-fried]/Kankee

Joseph moved to the Bay Area to work in AI research shortly after getting her undergraduate degree in neuroscience in 2019. There, she realized the social scene that seemed so sprawling online was far more tight-knit in person. Many rationalists and effective altruists, who call themselves EAs, worked together, invested in one another’s companies, lived in communal houses and socialized mainly with each other, sometimes in a web of polyamorous relationships. Throughout the community, almost everyone celebrated being, in some way, unconventional. Joseph found it all freeing and exciting, like winding up at a real-life rationalist Hogwarts. Together, she and her peers were working on the problems she found the most fascinating, with the rather grand aim of preventing human extinction. At the same time, she started to pick up weird vibes. One rationalist man introduced her to another as “perfect ratbait”—rat as in rationalist. She heard stories of sexual misconduct involving male leaders in the scene, but when she asked around, her peers waved the allegations off as minor character flaws unimportant when measured against the threat of an AI apocalypse. Eventually, she began dating an AI researcher in the community. She alleges that he committed sexual misconduct against her, and she filed a report with the San Francisco police. (Like many women in her position, she asked that the man not be named, to shield herself from possible retaliation.) Her allegations polarized the community, she says, and people questioned her mental health as a way to discredit her. Eventually she moved to Canada, where she’s continuing her work in AI and trying to foster a healthier research environment. “In an ideal world, the community would have had some serious discussions about sexual assault policy and education: ‘What are our blind spots? How could this have happened? How can we design mechanisms to prevent that from happening?’ ” she says. “I was disappointed how the community viewed me through this **very distorted**, **misogynistic** lens.” In Silicon Valley, the overlap between rationalists, EAs, and AI safety researchers forms a deeply influential subculture. While its borders are blurry, its hundreds or thousands of members are united by a belief that they need to work their butts off, or at least invest lots of money, to stop AI from going Terminator on us. The movement’s leaders have received support from some of the richest and most powerful people in tech, including Elon Musk, Peter Thiel and Ethereum creator Vitalik Buterin. And its ideas have attracted the usual Valley mix of true believers and brazen opportunists. Until recently, its most generous supporter was Bankman-Fried, who invested close to $600 million in related causes before dismissing effective altruism as a dodge once his business fell apart. Bankman-Fried’s collapse has cast a harsh light on the community’s flaws, but he’s far from the only alleged bad actor. The combination of insularity and shared purpose that makes the subculture so attractive to smart outsiders also makes it a hunting ground for con artists, sexual predators and megalomaniacs. Filtering the legitimate desire to make AI better and safer through the familiar lens of Valley messiah complexes risks tainting the whole project by association. The underlying ideology valorizes extremes: seeking rational truth above all else, donating the most money and doing the utmost good for the most important reason. This way of thinking can lend an attractive clarity, but it can also provide cover for destructive or despicable behavior. Eight women in these spaces allege pervasive sexual misconduct, including abuse and harassment, that they say has frequently been downplayed. Even among people with pure intentions, adherents say, EA and rationalist ideologies can amplify the suffering of people prone to doomsday thinking—leading, for a few, to psychotic breaks. These fissures have **global** **consequences**. The community’s connections and resources give its members outsize influence on the development of AI, the No. 1 object of fascination for today’s tech industry and an incredibly powerful tool worth untold billions. The believers are trying to make AI a force for good, but disillusioned members say their community of kindred spirits is being exploited and abused by people who don’t seem to know how to be humane. “Even if there’s a strong chance that bad AI outcomes will happen,” Joseph says, “using it as an excuse to erode human rights is disrespecting the very thing we’re fighting for.” The borders of any community this pedantic can be difficult to define. Some rationalists don’t consider themselves effective altruists, and vice versa. Many people who’ve drifted slightly from a particular orthodoxy hedge their precise beliefs with terms like “post-rationalist” or “EA-adjacent.” Yet two things are clear: Over the past decade EA has become the mainstream, public face of some fringe rationalist ideas, particularly the dire need for AI safety; and the whole thing started with Yudkowsky. Born in Chicago in 1979, Yudkowsky gave up Modern Orthodox Judaism as a preteen, becoming an atheist. He didn’t finish high school, but in his late teens he encountered and grew obsessed with the idea of the Singularity, the point at which technological progress will lead inevitably to superhuman intelligence. He started writing about AI in earnest in the 2000s, well after HAL 9000, Skynet and the Matrix had entered the public consciousness, but his prolificacy stood out. In years of pithy near-daily blog posts, he argued that researchers should do all they could to ensure AI was “aligned” with human values. To this end, he created the Machine Intelligence Research Institute (MIRI) in Berkeley, California, with early funding from Thiel and Skype co-founder Jaan Tallinn.

#### AI fears are irrational and exploited to advance authoritarian utilitarianism

Troy 23 [Dave Troy, investigative journalist from the Washington Spectator and Research Associate at Arizona State University educated at John Hopkins University, 5-1-2023, "The Wide Angle: Understanding TESCREAL — the Weird Ideologies Behind Silicon Valley’s Rightward Turn", Washington Spectator, https://washingtonspectator.org/understanding-tescreal-silicon-valleys-rightward-turn/]/Kankee

Lastly, Longtermism is a philosophy championed by MacAskill and his Oxford philosopher colleague Nick Bostrom. Mixing ideas from Russian Cosmism and E.A., Longtermism concerns itself with the maximization of future “intelligences” in the universe, and posits that anyone that interferes with that goal is harming countless future (potential) lives. This leads to some strange priorities, particularly a strong pro-natalist stance (you may recall that Musk has said that low birth rate is one of the biggest risks to humanity’s survival), but also a belief that in addition to biological intelligences, we should be maximizing machine intelligence in the universe. So that means not only should we be promoting biological space exploration and colonization (as per Cosmism), but we should also harness far-away planetary surfaces inhospitable to biologic life to build giant server “farms” from hypothetical materials like “computronium” — a kind of “programmable matter” that could host vast pools of mechanical Einsteins that could lead to the next big breakthroughs for intelligent life. If all of that sounds outlandish and orthogonal to solving the debt ceiling crisis, dealing with Earth’s climate problems, or otherwise improving conditions here on this planet, that’s because it is. TESCREAL proponents have an authoritarian “ends justify the means” mindset rooted in the idea that if we do not submit to their urgent demands, we will extinguish billions of potential future intelligent beings. Surely we must not allow that to happen! Eliezer Yudkowski, a self-described AI theorist, believes that AI is likely to wipe out humanity and that we should bomb data centers to stop its advance. Max Tegmark, an AI researcher at MIT, has also called for halting AI development in order to seek “alignment” — the idea that machine intelligence should work with humanity rather than against it. Such alarmist arguments, which originate in science fiction and are quite common in the TESCREAL world, are rooted in a hierarchical and zero-sum view of intelligence. The notion is that if we develop machine superintelligence, it may decide to wipe out less intelligent beings — like all of humanity. However, there is no empirical evidence to suggest these fears have any basis in reality. Some suggest that these arguments mirror ideas found in discredited movements like race science and Eugenics, even as others reject such charges. TESCREAL is a convergent Venn diagram of overlapping ideologies that, because they often attract contrarian young men, tend to co-occur with other male-dominated reactionary and misogynistic movements. The Men’s Rights movement (Manosphere), the MGTOW movement (Men Going Their Own Way), and PUA (Pick Up Artist) communities are near-adjacent to the TESCREAL milieu. Combining complex ideologies into such a “bundle” might seem to be dangerously reductive. However, as information warfare increasingly seeks to bifurcate the world into Eurasian vs. Atlanticist spheres, traditionalist vs. “woke,” fiat vs. hard currency, it’s difficult not to see the TESCREAL ideologies as integral to the Eurasianist worldview. I also independently identified these overlaps over the last few years, and thanks to philosopher Émile Torres and Dr. Gebru who together coined the TESCREAL acronym, we now have a shorthand for describing the phenomenon. As you encounter these ideologies in the wild, you might use the TESCREAL lens, and its alignment with Eurasianism and Putin’s agenda, to evaluate them, and ask whether they tend to undermine or enhance the project of liberal democracy. TESCREAL ideologies tend to advance an illiberal agenda and authoritarian tendencies, and it’s worth turning a very critical eye towards them, especially in cases where that’s demonstrably true. Clearly there are countless well-meaning people trying to use technology and reason to improve the world, but that should never come at the expense of democratic, inclusive, fair, patient, and just governance. The biggest risk AI poses right now is that alarmists will use the fears surrounding it as a cudgel to enact sweeping policy reforms. We should resist those efforts. Now more than ever, we should be guided by expertise, facts, and evidence as we seek to use technology in ways that benefit everyone.

#### AI existential risk focus distracts from non-hypothetical issues present AI companies perpetrate

Helfrich 24 [Gina Helfrich, Manager of the Centre for Technomoral Futures at Edinburgh Futures Institute at the University of Edinburgh and Deputy Chair of the University’s AI and Data Ethics Advisory Board with a Ph.D. in Philosophy from Emory University, 3-10-2024, "The harms of terminology: why we should reject so-called “frontier AI”", SpringerLink, https://link.springer.com/article/10.1007/s43681-024-00438-1]/Kankee

\*can be recut as a “frontier AI” word PIK

4 Defining “frontier AI” Frustratingly for our purposes, the definition of “frontier AI” is quite fuzzy. According to the UK government, “frontier AI” consists of “highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today’s most advanced models.” [18] Already with this definition, the emphasis is on the speculative future rather than the actual present. “Frontier AI” will “match or exceed” what “today’s most advanced models” are capable of.Footnote3 The term designates “highly capable foundation models that could possess dangerous capabilities.”Footnote4 In other words, “frontier AI” is speculative—it does not even exist yet, but the implication is that it is right around the corner. The idea of soon-to-be massively powerful AI technologies is a key indicator of the hype-inflated worldview inherent in the term “frontier AI.” The people popularizing this term are deeply concerned with what they call “existential risk” and “value alignment.” That is, they fear that a super-powerful AI could someday (perhaps quite soon) go rogue, leading to the destruction of humanity. As a result, they are invested in finding ways to attempt to align the “values” of AI with those of humanity. Often, the very same people making these claims are actually building the technologies they claim to so greatly fear.Footnote5 [19] Critics have argued that stoking fears of “existential risk” is really just another form of AI hype: if it is powerful enough to kill us all, then it must be really powerful. Focusing on far-future dangers is a means of diverting attention from the regulation-free status quo, which works in favor of the small cadre of companies developing these AI models. Not all AI experts [20] believe that “existential risk” is a reasonable fear, [21] but there are high-profile people in the field who do, and they have a lot of money to mobilize in service of promoting their views. In fact, the Effective Altruist movement is explicitly focused on earning as much money as possible (so that they can use those funds to “do the most good”). One notable example of how the EA philosophy can fail in spectacular ways is that former billionaire and convicted fraud [22] Sam Bankman-Fried was an EA adherent. Bankman-Fried was a high-profile member of the Effective Altruist community. As outlined above, the authors of the July paper introducing the term “frontier AI” have links to the Future of Humanity Institute at Oxford. FHI’s founding director is philosopher Nick Bostrom, who is well known for his interest in existential risk and other highly speculative potential futures with AI technologies. Bostrom is a leading figure in what has been dubbed the TESCREAL ideologies [23], a constellation of linked beliefs held by a small group of influential and controversial individuals who are attempting to steer the AI agenda toward the creation of artificial general intelligence (AGI). A significant portion of the concern around “existential risk” hangs on the fear that AGI will achieve self-consciousness and then turn its aims to the destruction of humanity. Given its origins and the connotations within its definition, anytime someone adopts the term “frontier AI,” they are effectively endorsing this very specific set of beliefs about AI and its capabilities. At the same time, using the term gives credence to the idea that this is a technology that already exists. If we constantly talk about “frontier AI” as though it is a real thing, the fact that the term apparently refers to the potential existence of certain powerful or dangerous AI capabilities becomes an inconvenient afterthought. 5 Connotation vs denotation The development of new vocabulary is often an exercise in trying to shift the discourse by using words that have not just a denotation—an explicit or direct meaning—but also a particular connotation—what the word implies or suggests. The words we choose to name things are often deeply connected to values and commitments that influence our perceptions of the thing that is named. “Pre-born baby” [24] and “embryo” or “fetus” nominally describe the same thing, but the former connotes a particular worldview and set of values, focused on opposition to abortion. People in the “pro-life” movementFootnote6 are eager to get others to adopt this vocabulary and framing. Similarly, the TESCREALists behind “frontier AI” are hoping to get this language adopted to advance their own agenda. So what might “frontier AI” connote? To start, the word “frontier” evokes the American “Wild West” and its colonial mentality: a place where civilized white men venture into the dangerous unknown and use lethal force to dominate and exploit what they find there, for profit. Others [25] have argued [26] that machine learning reproduces colonial logics, and the term “frontier AI” certainly invokes and reinscribes this colonial dynamic. We can also see the colonial dynamic between the handful of powerful Western companies who produce today’s generative AI models and the people of the “Global South” [27] who are most likely to experience harm as a direct result of the development and deployment of these AI technologies. The UK’s AI Safety Summit agenda apparently argued that the frontier is “where the risks are most urgent” and “where the vast promise of the future economy lies” [28]—marking the frontier as a place of great opportunity, but also great danger. Thus, “frontier AI” becomes a verbal proxy for the “existential risk” that TESCREALists are constantly pushing, with the added implication that such powerful technology can also lead to large profits for those able to harness it. A technological frontier also calls to mind the glittering expanse above us—space.Footnote7 TESCREALists like billionaire and SpaceX founder Elon Musk are keen to lead the charge into the stars, around which they plan to build vast computer simulations in which astronomical numbers of digital people will live out happy lives. [29] According to the flavor of utilitarianism known as longtermism (the “L” of “TESCREAL”), due to their overwhelming numbers, the sum total happiness of these hypothetical space-based future digital people dwarfs the well-being of all actually living people on Earth; therefore, ensuring that these future digital souls come into existence is, on their view, a moral duty. The path to bringing those souls into being starts with the development of AGI. The word “frontier” also implies a fixed boundary, slowly receding in a linear fashion,Footnote8 as technology makes steady progress into the future, wresting the known from the formerly unknown. This, too, is misleading. Technology does not progress in a steady, linear march. Sometimes it limps along, as in the “AI winter,” [30] until something changes to renew its progress, and sometimes people stand up to refuse a particular technology, halting its “progress” for the sake of values they see as more important. [31] Profit. Danger. Outer space. Progress. These are the connotations of “frontier AI.” It should be obvious that “frontier AI” is an exercise in AI hype, given these connotations. Today’s AI tools have not yet demonstrated that they have a sustainable business model. [32] The main dangers that “frontier AI” boosters are concerned with are hypothetical “existential risks.” We are nowhere near colonizing outer space. With generative AI threatening the livelihoods of creatives, writers, illustrators, and actors, it is clear that not everyone agrees that today’s AI tools represent progress. 6 Reasons to reject “frontier AI” We need to push back against the term “frontier AI” before it becomes uncritically adopted by the press and others. It is not clear that the thing TESCREALists want to name with the words “frontier AI” even exists—but if it does, then “highly capable generative AI models,” while less pithy, is perfectly serviceable and much more matter of fact. (I think even “foundation models” would be preferable.) As we have seen with terms like “foundation model” and “hallucination,” the press and others with an investment in staying up to date with technology are typically eager to adopt the latest terminology so as to appear in the know. (Special language is a way to distinguish insiders from outsiders, as anyone navigating the sea of acronyms at most universities can attest.) Pushing out the term “frontier AI” is a way for its boosters to both frame the conversation around what they think is important while simultaneously re-brandingFootnote9 large-scale generative machine learning models to, once again, divorce them from prior criticism. Large-scale generative machine learning models have been shown to cause several types of actual, non-hypothetical harms, including psychological harms [33, 34], social harms [35, 36], and environmental harms [37, 38]. “Frontier AI” encourages us to look past those very real harms in favor of a focus on the hypothetical future. Fortunately, the same researchers who have been leading the charge to uncover and highlight real AI harms and to develop robust AI governance mechanisms have already begun speaking out against efforts to shift the verbal terrain of AI. In a letter to the editor of the Financial Times responding to Hogarth’s “god-like AI” piece, Mhairi Aitken, Alan Turing Institute Ethics Fellow, wrote, “Words matter, and how we talk about AI has very real implications for how we engage with AI.” [39] Michael Birtwistle, the associate director of law and policy at the Ada Lovelace Institute, is quoted in the Guardian highlighting the hypothetical nature of “frontier AI:” “Policymaker attention and regulatory efforts are concentrated on a set of capabilities that don’t exist yet, a set of models that don’t yet show those capabilities.” [28] Meredith Whittaker, President of Signal, was quoted on the subject of “existential risk” as saying, “I think we need to recognize that what is being described, given that it has no basis in evidence, is much closer to an article of faith, a sort of religious fervor, than it is to scientific discourse.” [40] We should view the term “frontier AI” with skepticism if not outright suspicion, as yet another Trojan horse of AI hype. If we are to use this term at all, we should follow the standard set by Shannon Vallor and Ewa Luger, researchers and co-principal investigators of the UK government-funded “Bridging Responsible AI Divides” (BRAID) programme. [41] In a blog post excoriating the government’s sole focus on technical expertise as the path to “AI safety,” they correctly append the modifier “so-called” before “frontier AI” and always use the word “frontier” in scare quotes. [42] Journalists covering the current debates around AI should follow this practice; otherwise, they will effectively be endorsing the idiosyncratic views of those they ought to be reporting on with impartiality. Speaking of a hypothetical entity as though it is real puts journalists in the uncomfortable position of reporting on the future as though it has already happened. “Frontier AI” is AI hype. It is hypothetical, not real. It distracts focus away from AI’s actual harms by focusing on so-called “existential risk.” It carries with it connotations of colonialism and conquest that we should not be endorsing. The sooner everyone stops using this term, the better.

#### AGI fear mongering promises mass atrocities and permanent poverty for the sake of preventing existential risk

Robinson 21 [Émile P. Torres, postdoctoral researcher at Case Western Reserve University with a Bachelor of Science in philosophy and a Master of Science in neuroscience from Brandeis University, 7-28-2021, "The Dangerous Ideas of “Longtermism” and “Existential Risk”", Current Affairs, https://www.currentaffairs.org/news/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk]/Kankee

In the same paper, Bostrom declares that even “a non-existential disaster causing the breakdown of global civilization is, from the perspective of humanity as a whole, a potentially recoverable setback,” describing this as “a giant massacre for man, a small misstep for mankind.” That’s of course cold comfort for those in the crosshairs of climate change—the residents of the Maldives who will lose their homeland, the South Asians facing lethal heat waves above the 95-degree F wet-bulb threshold of survivability, and the 18 million people in Bangladesh who may be displaced by 2050. But, once again, when these losses are juxtaposed with the apparent immensity of our longterm “potential,” this suffering will hardly be a footnote to a footnote within humanity’s epic biography. These aren’t the only incendiary remarks from Bostrom, the Father of Longtermism. In a paper that founded one half of longtermist research program, he characterizes the most devastating disasters throughout human history, such as the two World Wars (including the Holocaust), Black Death, 1918 Spanish flu pandemic, major earthquakes, large volcanic eruptions, and so on, as “mere ripples” when viewed from “the perspective of humankind as a whole.” As he writes: “Tragic as such events are to the people immediately affected, in the big picture of things … even the worst of these catastrophes are mere ripples on the surface of the great sea of life.” In other words, 40 million civilian deaths during WWII was awful, we can all agree about that. But think about this in terms of the 1058 simulated people who could someday exist in computer simulations if we colonize space. It would require trillions and trillions and trillions of WWIIs one after another to even approach the loss of these unborn people if an existential catastrophe were to happen. This is the case even on the lower estimates of how many future people there could be. Take Greaves and MacAskill’s figure of 1018 expected biological and digital beings on Earth alone (meaning that we don’t colonize space). That’s still a way bigger number than 40 million—analogous to a single grain of sand next to Mount Everest. It’s this line of reasoning that leads Bostrom, Greaves, MacAskill, and others to argue that even the tiniest reductions in “existential risk” are morally equivalent to saving the lives of literally billions of living, breathing, actual people. For example, Bostrom writes that if there is “a mere 1 percent chance” that 1054 conscious beings (most living in computer simulations) come to exist in the future, then “we find that the expected value of reducing existential risk by a mere one billionth of one billionth of one percentage point is worth a hundred billion times as much as a billion human lives.” Greaves and MacAskill echo this idea in a 2021 paper by arguing that “even if there are ‘only’ 1014 lives to come … , a reduction in near-term risk of extinction by one millionth of one percentage point would be equivalent in value to a million lives saved.” To make this concrete, imagine Greaves and MacAskill in front of two buttons. If pushed, the first would save the lives of 1 million living, breathing, actual people. The second would increase the probability that 1014 currently unborn people come into existence in the far future by a teeny-tiny amount. Because, on their longtermist view, there is no fundamental moral difference between saving actual people and bringing new people into existence, these options are morally equivalent. In other words, they’d have to flip a coin to decide which button to push. (Would you? I certainly hope not.) In Bostrom’s example, the morally right thing is obviously to sacrifice billions of living human beings for the sake of even tinier reductions in existential risk, assuming a minuscule 1 percent chance of a larger future population: 1054 people. All of this is to say that even if billions of people were to perish in the coming climate catastrophe, so long as humanity survives with enough of civilization intact to fulfill its supposed “potential,” we shouldn’t be too concerned. In the grand scheme of things, non-runaway climate change will prove to be nothing more than a “mere ripple” —a “small misstep for mankind,” however terrible a “massacre for man” it might otherwise be. Even worse, since our resources for reducing existential risk are finite, Bostrom argues that we must not “fritter [them] away” on what he describes as “feel-good projects of suboptimal efficacy.” Such projects would include, on this account, not just saving people in the Global South—those most vulnerable, especially women—from the calamities of climate change, but all other non-existential philanthropic causes, too. As the Princeton philosopher Peter Singer writes about Bostrom in his 2015 book on Effective Altruism, “to refer to donating to help the global poor … as a ‘feel-good project’ on which resources are ‘frittered away’ is harsh language.” But it makes perfectly good sense within Bostrom’s longtermist framework, according to which “priority number one, two, three, and four should … be to reduce existential risk.” Everything else is smaller fish not worth frying. If this sounds appalling, it’s because it is appalling. By reducing morality to an abstract numbers game, and by declaring that what’s most important is fulfilling “our potential” by becoming simulated posthumans among the stars, longtermists not only trivialize past atrocities like WWII (and the Holocaust) but give themselves a “moral excuse” to dismiss or minimize comparable atrocities in the future. This is one reason that I’ve come to see longtermism as an immensely dangerous ideology. It is, indeed, akin to a secular religion built around the worship of “future value,” complete with its own “secularised doctrine of salvation,” as the Future of Humanity Institute historian Thomas Moynihan approvingly writes in his book X-Risk. The popularity of this religion among wealthy people in the West—especially the socioeconomic elite—makes sense because it tells them exactly what they want to hear: not only are you ethically excused from worrying too much about sub-existential threats like non-runaway climate change and global poverty, but you are actually a morally better person for focusing instead on more important things—risk that could permanently destroy “our potential” as a species of Earth-originating intelligent life. To drive home the point, consider an argument from the longtermist Nick Beckstead, who has overseen tens of millions of dollars in funding for the Future of Humanity Institute. Since shaping the far future “over the coming millions, billions, and trillions of years” is of “overwhelming importance,” he claims, we should actually care more about people in rich countries than poor countries. This comes from a 2013 PhD dissertation that Ord describes as “one of the best texts on existential risk,” and it’s cited on numerous Effective Altruist websites, including some hosted by the Centre for Effective Altruism, which shares office space in Oxford with the Future of Humanity Institute. The passage is worth quoting in full: “Saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive. By ordinary standards—at least by ordinary enlightened humanitarian standards—saving and improving lives in rich countries is about equally as important as saving and improving lives in poor countries, provided lives are improved by roughly comparable amounts. But it now seems more plausible to me that saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal.” Never mind the fact that many countries in the Global South are relatively poor precisely because of the long and sordid histories of Western colonialism, imperialism, exploitation, political meddling, pollution, and so on. What hangs in the balance is astronomical amounts of “value.” What shouldn’t we do to achieve this magnificent end? Why not prioritize lives in rich countries over those in poor countries, even if gross historical injustices remain inadequately addressed? Beckstead isn’t the only longtermist who’s explicitly endorsed this view, either. As Hilary Greaves states in a 2020 interview with Theron Pummer, who co-edited the book Effective Altruism with her, if one’s “aim is doing the most good, improving the world by the most that I can,” then although “there’s a clear place for transferring resources from the affluent Western world to the global poor … longtermist thought suggests that something else may be better still.” Returning to climate change once again, we can see how Tallinn got the idea that our environmental impact probably isn’t existentially risky from academic longtermists like Bostrom. As alluded to above, Bostrom maintains that non-runaway (which he calls “moderate”) global warming, as well as “threats to the biodiversity of Earth’s ecosphere,” as “endurable” rather than “terminal” for humanity. Similarly, Ord claims in The Precipice that climate change poses a mere 1-in-1,000 chance of existential catastrophe, in contrast to a far greater 1-in-10 chance of catastrophe involving superintelligent machines (dubbed the “Robopocalypse” by some). Although, like Bostrom, Ord acknowledges that the climate crisis could get very bad, he assures us that “the typical scenarios of climate change would not destroy our potential.” Within the billionaire world, these conclusions have been parroted by some of the most powerful men on the planet today (not just Tallinn). For example, Musk, an admirer of Bostrom’s who donated $10 million in 2015 to the Future of Life Institute, another longtermist organization that Tallinn cofounded, said in an interview this year that his “concern with the CO2 is not kind of where we are today or even … the current rate of carbon generation.” Rather, the worry is that “if carbon generation keeps accelerating and … if we’re complacent then I think … there’s some risk of sort of non-linear climate change”—meaning, one surmises, a runaway scenario. Peter Thiel has also apparently held this view for some time, which is unsurprising given his history with longtermist thinking and the Effective Altruism movement. (He gave the keynote address at the 2013 Effective Altruism Summit.) But Thiel also declared in 2014: “People are spending way too much time thinking about climate change” and “way too little thinking about AI.” The reference to AI, or “artificial intelligence,” here is important. Not only do many longtermists believe that superintelligent machines pose the greatest single hazard to human survival, but they seem convinced that if humanity were to create a “friendly” superintelligence whose goals are properly “aligned” with our “human goals,” then a new Utopian age of unprecedented security and flourishing would suddenly commence. This eschatological vision is sometimes associated with the “Singularity,” made famous by futurists like Ray Kurzweil, which critics have facetiously dubbed the “techno-rapture” or “rapture of the nerds” because of its obvious similarities to the Christian dispensationalist notion of the Rapture, when Jesus will swoop down to gather every believer on Earth and carry them back to heaven. As Bostrom writes in his Musk-endorsed book Superintelligence, not only would the various existential risks posed by nature, such as asteroid impacts and supervolcanic eruptions, “be virtually eliminated,” but a friendly superintelligence “would also eliminate or reduce many anthropogenic risks” like climate change. “One might believe,” he writes elsewhere, that “the new civilization would [thus] have vastly improved survival prospects since it would be guided by superintelligent foresight and planning.” Tallinn makes the same point during a Future of Life Institute podcast recorded this year. Whereas a runaway climate scenario is at best many decades away, if it could happen at all, Tallinn speculates that superintelligence will present “an existential risk in the next 10 or 50 years.” Thus, he says, “if you’re going to really get AI right [by making it ‘friendly’], it seems like all the other risks [that we might face] become much more manageable.” This is about as literal an interpretation of “deus ex machina” as one can get, and in my experience as someone who spent several months as a visiting scholar at the Centre for the Study of Existential Risk, which was cofounded by Tallinn, it’s a widely-held view among longtermists. In fact, Greaves and MacAskill estimate that every $100 spent on creating a “friendly” superintelligence would be morally equivalent to “saving one trillion [actual human] lives,” assuming that an additional 1024 people could come to exist in the far future. Hence, they point out that focusing on superintelligence gets you a way bigger bang for your buck than, say, preventing people who exist right now from contracting malaria by distributing mosquito nets. What I find most unsettling about the longtermist ideology isn’t just that it contains all the ingredients necessary for a genocidal catastrophe in the name of realizing astronomical amounts of far-future “value.” Nor is it that this religious ideology has already infiltrated the consciousness of powerful actors who could, for example, “save 41 [million] people at risk of starvation” but instead use their wealth to fly themselves to space. Even more chilling is that many people in the community believe that their mission to “protect” and “preserve” humanity’s “longterm potential” is so important that they have little tolerance for dissenters. These include critics who might suggest that longtermism is dangerous, or that it supports what Frances Lee Ansley calls white supremacy (given the implication, outlined and defended by Beckstead, that we should prioritize the lives of people in rich countries). When one believes that existential risk is the most important concept ever invented, as someone at the Future of Humanity Institute once told me, and that failing to realize “our potential” would not merely be wrong but a moral catastrophe of literally cosmic proportions, one will naturally be inclined to react strongly against those who criticize this sacred dogma. When you believe the stakes are that high, you may be quite willing to use extraordinary means to stop anyone who stands in your way. In fact, numerous people have come forward, both publicly and privately, over the past few years with stories of being intimidated, silenced, or “canceled.” (Yes, “cancel culture” is a real problem here.) I personally have had three colleagues back out of collaborations with me after I self-published a short critique of longtermism, not because they wanted to, but because they were pressured to do so from longtermists in the community. Others have expressed worries about the personal repercussions of openly criticizing Effective Altruism or the longtermist ideology. For example, the moral philosopher Simon Knutsson wrote a critique several years ago in which he notes, among other things, that Bostrom appears to have repeatedly misrepresented his academic achievements in claiming that, as he wrote on his website in 2006, “my performance as an undergraduate set a national record in Sweden.” (There is no evidence that this is true.) The point is that, after doing this, Knutsson reports that he became “concerned about his safety” given past efforts to censure certain ideas by longtermists with clout in the community. This might sound hyperbolic, but it’s consistent with a pattern of questionable behavior from leaders in the Effective Altruism movement more generally. For example, one of the first people to become an Effective Altruist after the movement was born circa 2009, Simon Jenkins, reports an incident in which he criticized an idea within Effective Altruism on a Facebook group run by the community. Within an hour, not only had his post been deleted but someone who works for the Centre for Effective Altruism actually called his personal phone to instruct him not to question the movement. “We can’t have people posting anything that suggests that Giving What We Can [an organization founded by Ord] is bad,” as Jenkins recalls. These are just a few of several dozen stories that people have shared with me after I went public with some of my own unnerving experiences. All of this is to say that I’m not especially optimistic about convincing longtermists that their obsession with our “vast and glorious” potential (quoting Ord again) could have profoundly harmful consequences if it were to guide actual policy in the world. As the Swedish scholar Olle Häggström has disquietingly noted, if political leaders were to take seriously the claim that saving billions of living, breathing, actual people today is morally equivalent to negligible reductions in existential risk, who knows what atrocities this might excuse? If the ends justify the means, and the “end” in this case is a veritable techno-Utopian playground full of 1058 simulated posthumans awash in “the pulsing ecstasy of love,” as Bostrom writes in his grandiloquent “Letter from Utopia,” would any means be off-limits? While some longtermists have recently suggested that there should be constraints on which actions we can take for the far future, others like Bostrom have literally argued that preemptive violence and even a global surveillance system should remain options for ensuring the realization of “our potential.” It’s not difficult to see how this way of thinking could have genocidally catastrophic consequences if political actors were to “[take] Bostrom’s argument to heart,” in Häggström’s words. I should emphasize that rejecting longtermism does not mean that one must reject long-term thinking. You ought to care equally about people no matter when they exist, whether today, next year, or in a couple billion years henceforth. If we shouldn’t discriminate against people based on their spatial distance from us, we shouldn’t discriminate against them based on their temporal distance, either. Many of the problems we face today, such as climate change, will have devastating consequences for future generations hundreds or thousands of years in the future. That should matter. We should be willing to make sacrifices for their wellbeing, just as we make sacrifices for those alive today by donating to charities that fight global poverty. But this does not mean that one must genuflect before the altar of “future value” or “our potential,” understood in techno-Utopian terms of colonizing space, becoming posthuman, subjugating the natural world, maximizing economic productivity, and creating massive computer simulations stuffed with 1045 digital beings (on Greaves and MacAskill’s estimate if we were to colonize the Milky Way). Care about the long term, I like to say, but don’t be a longtermist. Superintelligent machines aren’t going to save us, and climate change really should be one of our top global priorities, whether or not it prevents us from becoming simulated posthumans in cosmic computers. Although a handful of longtermists have recently written that the Effective Altruism movement should take climate change more seriously, among the main reasons given for doing so is that, to quote an employee at the Centre for Effective Altruism, “by failing to show a sufficient appreciation of the severity of climate change, EA may risk losing credibility and alienating potential effective altruists.” In other words, community members should talk more about climate change not because of moral considerations relating to climate justice, the harms it will cause to poor people, and so on, but for marketing reasons. It would be “bad for business” if the public were to associate a dismissive attitude about climate change with Effective Altruism and its longtermist offshoot. As the same author reiterates later on, “I agree [with Bostrom, Ord, etc.] that it is much more important to work on x-risk … , but I wonder whether we are alienating potential EAs by not grappling with this issue.” Yet even if longtermists were to come around to “caring” about climate change, this wouldn’t mean much if it were for the wrong reasons. Knutsson says: “Like politicians, one cannot simply and naively assume that these people are being honest about their views, wishes, and what they would do. In the Effective Altruism and existential risk areas, some people seem super-strategic and willing to say whatever will achieve their goals, regardless of whether they believe the claims they make—even more so than in my experience of party politics.” Either way, the damage may already have been done, given that averting “untold suffering” from climate change will require immediate action from the Global North. Meanwhile, millionaires and billionaires under the influence of longtermist thinking are focused instead on superintelligent machines that they believe will magically solve the mess that, in large part, they themselves have created.

#### Reject AI threat surveys – its AI doomerism funded by EA for more funding

**Stokel-Walker 24** [Chris Stokel-Walker, British freelance journalist and Bachelor of the Arts in English Literature at Newcastle University, 2024, "AI Survey Exaggerates Apocalyptic Risks," Scientific American, https://www.scientificamerican.com/article/ai-survey-exaggerates-apocalyptic-risks/]/Kankee

That was the sobering finding of a paper posted on the preprint server arXiv.org. In it, the authors reported the results of a survey of 2,778 researchers who had presented and published work at high-profile AI research conferences and journals—the biggest such poll to date in a once-obscure field that has suddenly found itself navigating core issues of humanity’s future. “People are interested in what AI researchers think about these things,” says Katja Grace, co-lead author of the paper and lead researcher at AI Impacts, the organization that conducted the survey. “They have an important role in the conversation about what happens with AI.” But some AI researchers say they’re concerned the survey results were biased toward an alarmist perspective. AI Impacts has been partially funded by several organizations, such as Open Philanthropy, that promote effective altruism—an emerging philosophical movement that is popular in Silicon Valley and known for its doom-laden outlook on AI’s future interactions with humanity. These funding links, along with the framing of questions within the survey, have led some AI researchers to speak up about the limitations of using speculative poll results to evaluate AI’s true threat. Effective altruism, or EA, is presented by its backers as an “intellectual project” aimed at using resources for the greatest possible benefit to human lives. The movement has increasingly focused on AI as one of humanity’s existential threats, on par with nuclear weapons. But critics say this preoccupation with **speculative** future scenarios **distracts** society from the discussion, research and regulation of the risks AI already poses today—including those involving discrimination, privacy and labor rights, among other pressing problems. The recent survey, AI Impacts’ third such poll of the field since 2016, asked researchers to estimate the probability of AI causing the “extinction” of humanity (or “similarly permanent and severe disempowerment” of the species). Half of respondents predicted a probability of 5 percent or more. But framing survey queries this way inherently promotes the idea that AI poses an existential threat, argues Thomas G. Dietterich, former president of the Association for the Advancement of Artificial Intelligence (AAAI). Dietterich was one of about 20,000 researchers who were asked to take part—but after he read through the questions, he declined. “As in previous years, many of the questions are asked from the AI-doomer, existential-risk perspective,” he says. In particular, some of the survey’s questions directly asked respondents to assume that high-level machine intelligence, which it defined as a machine able to outperform a human on every possible task, will eventually be built. And that’s not something every AI researcher sees as a given, Dietterich notes. For these questions, he says, almost **any** result could be used to support **alarming** conclusions about AI’s potential future. “I liked some of the questions in this survey,” Dietterich says. “But I still think the focus is on ‘How much should we worry?’ rather than on doing a careful risk analysis and setting policy to mitigate the relevant risks.” Others, such as machine-learning researcher Tim van Erven of the University of Amsterdam, took part in the survey but later regretted it. “The survey emphasizes baseless speculation about human extinction without specifying by which mechanism” this would happen, van Erven says. The scenarios presented to respondents are not clear about the hypothetical AI’s capabilities or when they would be achieved, he says. “Such vague, hyped-up notions are dangerous because they are being used as a smokescreen ... to draw attention away from mundane but much more urgent issues that are happening right now,” van Erven adds. Grace, the AI Impacts lead researcher, counters that it’s important to know if most of the surveyed AI researchers believe existential risk is a concern. That information should “not necessarily [be obtained] to the exclusion of all else, but I do think that should definitely have at least one survey,” she says. “The different concerns all add together as an emphasis to be careful about these things.” The fact that AI Impacts has received funding from an organization called Effective Altruism Funds, along with other backers of EA that have previously supported campaigns on AI's existential risks, has prompted some researchers to suggest the survey’s framing of existential-risk questions may be influenced by the movement. Nirit Weiss-Blatt, a communications researcher and journalist who has studied effective altruists’ efforts to raise awareness of AI safety concerns, says some in the AI community are uncomfortable with the focus on existential risk—which they claim comes at the expense of other issues. “Nowadays, more and more people are reconsidering letting effective altruism set the agenda for the AI industry and the upcoming AI regulation,” she says. “EA’s reputation is deteriorating, and backlash is coming.” “I guess to the extent that criticism is that we are EAs, it’s probably hard to head off,” Grace says. “I guess I could probably denounce EA or something. But as far as bias about the topics, I think I’ve written none of the best pieces on the counterarguments against thinking AI will drive humanity extinct.” Grace points out that she herself doesn’t know all her colleagues’ beliefs about AI’s existential risks. “I think AI Impacts overall is, in terms of beliefs, more all over the place than people think,” she says. Defending their research, Grace and her colleagues say they have worked hard to address some of the criticisms levelled at AI Impacts’ studies from previous years—especially the argument that relatively low numbers of respondents hadn’t adequately represent the field. This year the AI Impacts team tried to boost the number of respondents by reaching out to more people and expanding the conferences from which it drew participants. But some say this dragnet still isn’t wide enough. “I see they’re still not including conferences that think about ethics and AI explicitly, like FAccT [the Association for Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency] or AIES [the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society],” says Margaret Mitchell, chief ethics scientist at AI company Hugging Face. “These are the ‘top AI venues’ for AI and ethics.” Mitchell received an invitation to join the survey but didn’t do so. “I generally just don't respond to e-mails from people I don't know asking me to do more work,” she says. She speculates that this kind of situation could help skew survey results. “You're more likely to get people who don't have tons of e-mail to respond to or people who are keen to have their voices heard—so more junior people,” she says. “This may affect hard-to-quantify things like the amount of wisdom captured in the choices that are made.” But there is also the question of whether a survey asking researchers to make guesses about a far-flung future provides any valuable information about the ground truth of AI risk at all. “I don’t think most of the people answering these surveys are[n’t] performing a careful risk analysis,” Dietterich says. **Nor** are they asked to back up their predictions. “If we want to find useful answers to these questions,” he says, “we need to fund research to carefully assess each risk and benefit.”

#### Big tech securitizes AI to ensure regulatory capture and stifle competition

**Davidson 23** [John Davidson, Columnist at Australian Financial Review and Grad Cert in Creative Writing at University of Technology Sydney, 2023, "Google Brain founder says big tech is lying about AI extinction danger," Australian Financial Review, https://www.afr.com/technology/google-brain-founder-says-big-tech-is-lying-about-ai-human-extinction-danger-20231027-p5efnz]/Kankee

The notion that artificial intelligence could lead to the extinction of humanity is a “bad idea” being **promulgated** by big tech in the hope of triggering heavy regulation that would **shut down** competition in the AI market, one of the world’s top AI experts warned. Andrew Ng, a professor at Stanford University who taught machine learning to the likes of OpenAI co-founder Sam Altman, and who himself co-founded Google Brain and was chief scientist at Baidu’s Artificial Intelligence Group, said that the “bad idea that AI could make us go extinct” was merging with the “bad idea that a good way to make AI safer is to impose burdensome licensing requirements” on the AI industry. “When you put those two bad ideas together, you get the massively, colossally dumb idea [of] policy proposals that try to require licensing of AI,” Professor Ng told The Australian Financial Review in an interview. “It would crush innovation,” he said. “There are definitely large tech companies that would rather not have to try to compete with open source [AI], so they’re creating fear of AI leading to human extinction. “It’s been a **weapon** for lobbyists to argue for legislation that would be **very damaging** to the open-source community,” he said. In May, OpenAI CEO and co-founder Altman co-signed a letter saying that “mitigating the risk of extinction from AI should be a global priority”, and in March, more than 1100 industry leaders including Elon Musk and Apple co-founder Steve Wozniak called for a six-month moratorium on training powerful AI models. “Sam [Altman] was one of my students at Stanford. He interned with me. I don’t want to talk about him specifically because I can’t read his mind, but …I feel like there are many large companies that would find it convenient to not have to compete with open-sourced large language models,” he said. “There’s a standard regulatory capture playbook that has played out in other industries, and I would hate to see that executed successfully in AI.” Professor Ng declined to comment on the risk-based regulation of AI being proposed by the Labor government, but agreed that AI should be regulated. “I don’t think no regulation is the right answer, but with the direction regulation is headed in a lot of countries, I think we’d be better off with no regulation than what we’re getting,” he said. “But thoughtful regulation would be much better than no regulation,” he said. “Just to be clear, AI has caused harm. Self-driving cars have killed people. In 2010, an automated trading algorithm crashed the stock market. Regulation has a role. But just because regulation could be helpful doesn’t mean we want bad regulation.” High on the list of “good” regulations, he said, was the need for transparency from technology companies, which would have helped avert the social media disaster caused by big tech at the start of the century, and which will help avert AI disasters caused by big tech in the future.

#### AI fearmongering empowers authoritarian government takeovers, destroying democracy

**Chilson 24** [Neil Chilson, head of AI policy at the Abundance Institute and J.D. in law at The George Washington University Law School, 7-5-2024, "The authoritarian side of effective altruism comes for AI", Reason, https://reason.com/2024/07/05/the-authoritarian-side-of-effective-altruism-comes-for-ai/]/Kankee

The effective altruism (E.A.) movement, which began with the premise that philanthropists should do the most good per dollar spent, injected pragmatism into an arena where good intentions can trump rational, effective number crunching. A generation of effective altruists—many schooled in Silicon Valley thought—have since embraced this metric-driven, impartial philosophy and translated their own good intentions into good works. However, artificial intelligence (AI) exposes a flaw in the movement: a powerful faction of doomsayers. The result is not just misplaced philanthropy but lobbying to create agencies with utterly alarming authority. For various reasons, the E.A. movement has turned its attention toward longtermism—a more radical form of its utilitarianism that weighs the value of each future potential life approximately the same as a living person's. Because any human extinction event, however unlikely, imposes infinite costs, longtermists can place enormous moral value on reducing whatever they view as existential risk. Certain proponents of E.A. argue that intelligent-enough AIs pose such risk. Indeed, one of the most influential and longest-standing E.A. organizations, the Machine Intelligence Research Institute (MIRI), recently stated that its "objective is to convince major powers to shut down the development of frontier AI systems worldwide." MIRI's founder, Eliezer Yudkowsky, notoriously called on the U.S. to **bomb** "rogue" data centers and **threaten nuclear war** against countries that don't halt AI research. Extremism is not unique to AI debates. Environmentalists have Just Stop Oil and their unadvisable displays, religions have violent extremists, and even Luddites have the Unabomber. But in E.A. the radicals are **prominent**. Sam Bankman-Fried claimed his cryptocurrency scams were his Machiavellian plots to supply tens of millions of dollars to E.A. organizations, including through his own Future Fund. Despite blots on their reputation, AI doomers have hundreds of millions of E.A. dollars in backing. And while extremists publish manifestos, they rarely propose legislation exposing just how far they're willing to go—until now. Enter two proposed bills: the federal Responsible Advanced Artificial Intelligence Act (RAAIA) drafted by the Center for AI Policy, and California's Senate Bill 1047 sponsored by the Center for AI Safety (CAIS). Both bills and their backers are closely tied to E.A. and longtermist funding and organizations. The RAAIA is, simply put, **shocking** in its **authoritarianism**. The bill would create a new federal agency (run by an administrator appointed by the president) to govern a wide range of AI systems, from weather forecasting to weapons. Companies must get permits before developing software, which the agency can arbitrarily condition. If a permitted model proves too competent, the agency can halt the research. Open-source projects must somehow verify and track the identities of all users and ensure each has a "legitimate, pro-social interest." The emergency powers the RAAIA would grant to the president and administrator are **dictatorial**. The administrator can, on his own authority, shut down the entire frontier AI industry for six months. If the president declares an AI emergency, the administrator can seize and destroy hardware and software, enforced by guards "physically removing any unauthorized persons from specified facilities" and/or "taking full possession and control of specified locations or equipment." They can conscript the FBI and federal Marshals and direct other federal law enforcement officers. The administrator would have the prosecutorial and enforcement powers to subpoena witnesses, compel testimony, conduct raids, and demand any evidence deemed relevant, even for speculative "proactive" investigations. Further, the RAAIA would create a registry for all high-performance AI hardware. If you "buy, sell, gift, receive, trade, or transport" even one covered microchip without the required form, you will have committed a crime. The bill imposes criminal liability for other violations, and agency employees can be criminally prosecuted for "willfully and intentionally" refusing to perform duties prescribed by the act. The bill also includes tricks attempting to insulate the administrator from influence by future administrations or other aspects of government. For example, there's a one-way ratchet clause empowering the administrator to update rules but making it difficult to "weaken or loosen'' them. It attempts to constrain the judicial standard of review, compress appeal timeframes, and exempt the administrator from the Congressional Review Act, among other things. Predictably, the agency is funded through its imposed fines and fees. This creates an incentive to levy them, limits congressional budgetary oversight, and demonstrates the backers' disdain for democratic checks and balances. While the language in California's S.B. 1047 is milder, CAIS and state Rep. Scott Wiener (D–San Francisco) have written a state bill that could have a similarly authoritarian effect. S.B. 1047 would impose a new Frontier Model Division (FMD) to regulate organizations training AI models that require more than a certain threshold of computer power or expense—a threshold the FMD would set. Cloud computer providers would be required to implement a kill switch to shut down AI models if anything goes wrong, and additional emergency authorities would be given to the governor. But at its core, S.B. 1047 requires AI developers to prove a negative to a hostile regulator before proceeding. Specifically, developers of certain high-cost models must—somehow—prove ahead of time that their product could never be used to cause "critical damages." Variations of the word reasonable appear over 30 times in S.B. 1047. Of course, the FMD determines how reasonable is defined. Other weasel words used include material, good faith, and reasonably foreseeable. Wiener and his co-authors have hidden their authoritarianism in this vague and arbitrary language. If the FMD—likely staffed with E.A.-influenced AI doomers like those who wrote the bill—doesn't like an AI research proposal, it can impose custom conditions or block it entirely. Even if the FMD approves a plan, it can later determine that the plan was unreasonable and punish the company. All of this will inevitably deter the development of new models, which is perhaps the point. The deceptively milder language of S.B. 1047 is partly why it has already passed the California state Senate and is moving through the House. For now, the RAAIA lacks congressional sponsorship. Yet both bills should merit alarm. They are products of a radical E.A. faction that, in its fervor to regulate away a perceived threat, is willing to **blindly** empower governments through **unaccountable** agencies, vague requirements, presumption of guilt, and unchecked emergency powers.

#### Silicon Valley AI scaremongering and threat construction allows their capitalistic consolidation of power, obscuring killing machines, mass surveillance, and digital totalitarianism

**Ongweso 23** [Edward Ongweso Jr., senior researcher at Security in Context and BA in Politics, Philosophy, and Economics at Hampshire College, 06-07-2023, "AI Doesn’t Pose an Existential Risk—but Silicon Valley Does", Nation, https://www.thenation.com/article/economy/artificial-intelligence-silicon-valley/]/Kankee

A coalition of the willing has united to confront what they say is a menace that could destroy us all: artificial intelligence. More than 350 executives, engineers, and researchers who work on AI have signed a pithy one-sentence statement: “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks, such as pandemics and nuclear war.” But like the target of the last infamous coalition of the willing—Saddam Hussein and his mythical “weapons of mass destruction”—there is no existential threat here. This isn’t the first letter to sound the alarm. It features prominent figures in the field—such as Sam Altman, chief executive of Microsoft-backed OpenAI. Generally, the warnings about AI are straightforward: It poses immediate risks like discrimination or automation as well as existential ones like a superintelligent Skynet-like system eradicating humanity. These claims of an extinction-level threat come from the very same groups creating the technology, and their warning cries about future dangers is **drowning** **out** stories on the harms already occurring. There is an **abundance** of research documenting how AI systems are being used to steal art, control workers, expand private surveillance, and seek greater profits by replacing workforces with algorithms and underpaid workers in the Global South. The sleight-of-hand trick shifting the debate to existential threats is a marketing strategy, as Los Angeles Times technology columnist Brian Merchant has pointed out. This is an attempt to generate interest in certain products, dictate the terms of regulation, and protect incumbents as they develop more products or further integrate AI into existing ones. After all, if AI is really so dangerous, then why did Altman threaten to pull OpenAI out of the European Union if it moved ahead with regulation? And why, in the same breath, did Altman propose a system that just so happens to protect incumbents: Only tech firms with enough resources to invest in AI safety should be allowed to develop AI. No, the **real** **threat** is the industry that controls our technology ecosystem and lobbies for insulation from states and markets that might rein it in. I want to talk about three factors that make Silicon Valley, not one of its many developments, a “societal-scale risk.” First, the industry represents the culmination of various lines of thought that are deeply hostile to democracy. Silicon Valley owes its existence to state intervention and subsidy, at different times working to capture various institutions or wither their ability to interfere with private control of computation. Firms like Facebook, for example, have argued that they are not only too large or complex to break up but that their size must actually be protected and integrated into a geopolitical rivalry with China. Second, that hostility to democracy, more than a singular product like AI, is amplified by profit-seeking behavior that constructs increasingly larger threats to humanity. It’s Silicon Valley and its emulators worldwide, not AI, that create and finance harmful technologies aimed at surveilling, controlling, exploiting, and killing human beings with little to no room for the public to object. The search for profits and excessive returns, with state subsidy and intervention clearing the way of competition, has and will create a litany of immoral business models and empower **brutal** regimes alongside “existential” threats. At home, this may look like the surveillance firm and government contractor Palantir creating a deportation machine that terrorizes migrants. Abroad, this may look like the Israeli apartheid state exporting spyware and weapons it has tested on Palestinians. Third, this combination of a deeply antidemocratic ethos and a desire to seek profits while externalizing costs can’t simply be regulated out of Silicon Valley. These are fundamental attributes of the industry that trace back to the beginning of computation. These origins in optimizing plantations and crushing worker uprisings prefigure the obsession with surveillance and social control that shape what we are told technological innovations are for. Taken altogether, why should we worry about some far-flung threat of a superintelligent AI when its creators—an insular network of libertarians building digital plantations, surveillance platforms, and killing machines—exist here and now? Their Smaugian hoards, their fundamentalist beliefs about markets and states and democracy, and their track record should be impossible to ignore. Despite the constant crowing about how integral technology is to our society, you and I play virtually no role in deciding what gets built, who builds it, how it gets financed, or why it should be built. The small role the public plays largely boils down to ratification through channels that are built to accommodate larger vessels—states, markets, trade blocs, corporations, capital, political party institutions, robust lobbying networks complete with friends and insiders, and more. Powerful participants formulate policy in private and say “do so” to a public that’s actively excluded. Contempt for democracy is nothing new, of course. In America, it’s a vaunted pastime that stretches back to the start of our grand experiment. In debates at the Constitutional Convention, James Madison was unambiguous that their government’s goal was “to protect the minority of the opulent against the majority.” The Senate, he argued, would be instrumental to this purpose because that purpose would ensure the creation of “a system which we wish to last for ages.” Still, Madison argued, there was a key tension everyone was overlooking: An increase of population will of necessity increase the proportion of those who will labour under all the hardships of life, & secretly sigh for a more equal distribution of its blessings. These may in time outnumber those who are placed above the feelings of indigence. According to the equal laws of suffrage, the power will slide into the hands of the former. No agrarian attempts have yet been made in this Country, but symtoms [sic], of a leveling spirit, as we have understood, have sufficiently appeared in a certain quarters to give notice of the future danger. Peter Thiel—the billionaire cofounder of surveillance firm Palantir, head of the VC firm Founders Firm, and former board member of Facebook—has lamented similar outcomes. In a 2009 essay for Cato Unbound, Thiel admitted, “I no longer believe that freedom and democracy are compatible.” While the wake of the 2008 financial crisis affirmed to him and his fellow libertarians that “the broader education of the body politic has become a fool’s errand,” Thiel believed the problem went back further: “The roaring 1920s were the last decade where one could be genuinely optimistic about politics but since then we’ve seen a troubling development: Since 1920, the vast increase in welfare beneficiaries and the extension of the franchise to women—two constituencies that are notoriously tough for libertarians—have rendered the notion of ‘capitalist democracy’ into an oxymoron.” Thiel later clarified that he did not believe that disenfranchising women, or any other group, was desirable. He was simply saying suffrage posed a danger to other rights. There’s a clear thread here: Democracy is a virtue to pay lip service to, but there are other more important priorities that, if left on their own, the public will bungle. Such as politics. In the 20th century, American liberals who were concerned about the public’s ability to interfere in political affairs took up the thorny question of how elites should ensure control of America’s unruly democracy. Edward Bernays—Sigmud Freud’s nephew and the “father of spin”—argued in his 1928 book Propaganda that “conscious and intelligent manipulation of the organized habits and opinions of the masses is an important element of democratic society.” Why? Because “intelligent minorities” in our society need to “make use of propaganda continuously and systematically.” At around the same time, Walter Lippmann, a prominent and influential journalist whose major works maintained that reality was becoming too complex for the masses to understand, argues that an inability to distinguish reality from opinion necessitated “the manufacture of consent” to ensure that democracy functioned as desired. A “specialized class” of individuals with the foresight and position to realize those interests would manage those unable to. Among this specialized class, you’d have “public men,” who could ensure “the formation of a sound public opinion.” Lippman wanted to keep the public far from the formation, deliberation, and execution of affairs concerning them. “Public opinion is not a rational force,” Lippman wrote. “It does not reason, investigate, invent, persuade, bargain or settle.” The indolent public’s purpose is to ratify things already deliberated on. The democratic question isn’t something only liberals have been wrestling with, however. In his recent book Crack Up Capitalism, economic historian Quinn Slobodian documents the intellectual history and consequences of the capitalist right’s attempt to liberate capitalism from democracy. Largely drawn from the superrich, these libertarian utopians are searching for the ideal container for capitalism, for zones of exception where holes can be punched into nation-states to undermine the capacity for democracy to interfere with markets. “Champions of the zone suggest that free-market utopia might be reached through acts of secession and fragmentation, carving out liberated territory within and beyond nations, with both disciplining and demonstration effects,” Slobodian writes in the opening pages of his book. The text is littered with examples that span the globe. The book’s case studies pour over the obsession with city-state dictatorships like Dubai, Singapore, and Hong Kong. In one chapter, libertarians spend the 1980s trying to save Ciskei, a South African Bantustan, by forming a commission to explore how best to become “African Hong Kong.” The goal was not to eradicate apartheid but engineering a scenario “inviting in foreign capital while encouraging voluntary segregation from below instead of mandatory segregation from above.” Heading the commission was Leon Louw, a libertarian Afrikaner who founded the Free Market Foundation and styled himself as an abolitionist who could liberate the market from the apartheid democracy. Foreign capital came, not just for the state subsidies but also because of Ciseki’s eagerness to use force on the population. Workers were regularly detained and tortured; the police killed protesters; activists were assassinated; but investors’ needs were satisfied. There are other examples closer to Silicon Valley: Saudi Arabia’s delusional NEOM, former Andreessen Horowitz (a16z) partner Balaji Srinivasan’s grand strategy to put nation-states on the cloud, and Thiel’s dream for similar vision for an escape beyond politics—a retreat into colonizing outer space, cyberspace, and the oceans. There was also the half-baked plan pitched by Stanford economics professor Paul Romer to craft what reactionary blogger Curtis Yarvin called a “colonialism for the 21st century” and apply it to Honduras. A plan, Slobodian points out, that had commentators drooling over its vision and ambition and slick, forward-looking momentum. Peter Thiel ended that 2009 essay with a sweet note saying that “all of us must wish Patri Friedman the very best in his extraordinary experiment.” That experiment was to carve out Romer’s enclave in Honduras with the help of people in Thiel’s orbit, backed by investors from the Future Cities Development group (which Friedman cofounded), and bring the “Silicon Valley spirit of innovation to Honduras.” Other investors came; memorandums and agreements with the government were signed; ideologues spoke about the potential of this experiment to revolutionize sovereignty and governance. Many of these people were also attracted to a project called Prospera, which was built on an island off the coast of Honduras. Prospera was not only built and funded by networks involved in Romer’s adventure but also managed to extract a territorial concession from Honduras and lobby for a law that allowed corporations to set up zones in the country. “While earlier settlers once sought wealth in gold, crops, or railroads, the treasure of zones like Prospera in the twenty-first century was their status as a jurisdiction—their potential as a new place to pick and choose among regulations and licensing requirements,” Slobodian explained. Crucially, the anarcho-capitalists who inspired and helped make this project sought to make a colony where the social contract was “a literal contract” shaped by whatever regulations investors were interested in adhering to or skirting. That sort of perforation was helped along by the fact that Honduras was, like Ciseki, liberal with its use of force. The Honduran government had already spent decades detaining, torturing, and murdering protesters and activists. For libertarians and their utopias, this willingness to use violence against a population that might protest is important, but it is not sufficient for their control. Ciskei collapsed, and Prospera may soon follow. Last year, the Honduran government rejected the law and constitutional amendment enabling Zones for Employment and Economic Development, the corporate enclaves that libertarians have been so excited about. Proponents of the liberal variant of antidemocratic thought are also concerned with ensuring that the public and the instruments responsive to it—like the state—don’t get in the way of their own self interest. As I wrote in my previous article, post-WWII planners and their Silicon Valley tech heirs insist that we can solve various crises (i.e., ecological catastrophe or permanent surveillance systems) only by handing over control to them—the very saboteurs responsible for these crises. Among tech elites, sometimes the general principle that specialized classes alone have the education, position, and inherent ability to act calmly and rationally based on the facts is said loudly. In mid-May, former Google chief executive Eric Schmidt told NBC’s Meet the Press that Big Tech and Big Tech alone should regulate artificial intelligence. “When this technology becomes more broadly available, which it will, and very quickly, the problem will get worse. I would much rather have the current companies define reasonable boundaries,” Schmidt said in the interview. “There’s no way a non-industry person can understand what’s possible. It’s just too new, too hard; there’s not the expertise. There’s no one in the government that can get it right. The industry can broadly get it right.” Schmidt has spent years ringing the alarm bell about artificial intelligence, arguing that it will be a key geopolitical fault line and that we risk ceding it to China. In an essay for Le Monde, tech critic Evgeny Morozov dives a bit deeper into Schmidt and connects him to Gilman Louie—a key figure in the coming US-China cold war (Cold War 2.0) who worked with the Air Force, ran the CIA’s venture capital fund, worked with Schmidt at the National Security Commission on Artificial Intelligence, and now runs the Schmidt-backed America’s Frontier Fund. “Ironically, Gilman Louie, the man who leveraged Cold War 1.0 to hype up Tetris, is now leveraging Cold War 2.0 to hype up AI. Or perhaps vice versa,” Morozov writes. “In today’s Washington, these two operations have become almost indistinguishable, and the only certainty is that all that hype will be monetised.” Scaremongering about AI is a tactic to sell more AI. But it’s also part of a larger campaign that poses an actual threat to all of us. A deeply entrenched **contempt** for democracy, a desire to use the state as a vessel for reshaping society into something more amenable to unregulated development and profit-seeking, and a long-standing obsession with surveillance and social control will deliver eye-watering returns for a few. It will also leave us with a world **dominated** by innovative extraction, violent borders, robust and dynamic repression, and streamlined **violence**. Don’t fall for the trick: Silicon Valley, not AI, is the **existential** risk to humanity.

#### Longtermism destroys the moral meaning of present racialized suffering and guarantees long-term oppression

**Crary 23** [Alice Crary, University Distinguished Professor at the New School for Social Research and Ph.D. in philosophy at University of Pittsburgh, 2023, "The toxic ideology of longtermism", Radical Philosophy, https://www.radicalphilosophy.com/commentary/the-toxic-ideology-of-longtermism]/Kankee

Longtermism’s sins are different and more ominous, but there are points of convergence. Longtermism deflects from EA’s wonted attention to current human and animal suffering. It defends in its place a concern for the wellbeing of the potentially trillions of humans who will live in the long-term future, and, taking the sheer number of prospective people to drown out current moral problems, exhorts us to regard threats to humanity’s continuation as a moral priority, if not the moral priority.11 This makes longtermists shockingly dismissive of ‘nonexistential’ hazards that may result in the suffering and death of huge numbers in the short term if, as they see it, there is a reasonable probability that the hazards are consistent with the possibility of a far greater number of humans going on to flourish in the long term. When longtermists turn to existential hazards, they discuss wholly natural threats (such as large asteroids hurtling toward the earth, super-volcanic eruptions, and stellar explosions) while focusing on human caused risks, which they regard as more likely to rise to extinction-level. Alongside value-divergent AI and human-produced pathogens, they consider climate change, other forms of environmental degradation, and all-out nuclear war, and they set out to calculate the probability that these different anthropogenic threats will instigate existential disasters. This accent on existential dangers is theoretically unjustified and morally damaging, but even stripped of it, longtermism is a poor guide to solicitude for prospective humans. Longtermism calls on us to safeguard humanity’s future in a manner that both diverts attention from current misery and leaves harmful socioeconomic structures critically unexamined. As a movement, it has enjoyed stunning financial success and clout. But its success is not due to the quality of its conception of morality, which builds questionably on EA’s. Rather, it is due to longtermism’s compatibility with the very socioeconomic arrangements that have led us to the **brink** of the kinds of catastrophes it claims to be staving off. At issue is not only an especially dangerous, future-facing variation on ideologies, like EA, that **thwart** struggles for liberating change with suggestions of the cure-all properties of existing economic tools. It is a variation lacking any plausible rationale, since many of these struggles have long contributed to the area longtermism wrongly represents as its innovation – fighting for a just and livable future. The longtermist enterprise has been publicly thrashed for its ties to FTX, but it remains well-funded and well-positioned to repair its reputation and go on enlisting earnest individuals to energetically support and spread it. There is a pressing need to criticise its theoretical weaknesses and forcefully bring out its material harms, exposing it as the **toxic** ideology it is. Longtermist moral logic The ethical core of longtermism is a set of commitments, shared with EA, from the moral tradition of consequentialism. For consequentialists, the mark of right action is producing outcomes that are best in the sense of containing the greatest amount of value. That leaves open what is of value, and, although longtermists often insist on respect for uncertainty about the correctness of any one moral theory, they still incline toward versions of consequentialism that identify value with wellbeing and so fall under the heading of utilitarianism. In making these theoretical moves, longtermists help themselves to a methodological assumption that is itself morally significant. Together with effective altruists and many others partial to utilitarian stances, they assume that wellbeing is discernible from a dispassionate and abstract ‘point of view of the universe’.12 That is morally significant because it is what seems to make it possible to use wellbeing as a measure for comparing outcomes anywhere – not only across space to the global poor and across species to non-human animals but also across time to those living in the far distant future. Longtermism proper emerges from within a set of contemporary ethical discussions, typically described as composing the field of ‘population ethics’, in which utilitarianism-tinged modes of thought are applied to prospective humans.13 Debates among population ethicists pivot around questions about whether our moral assessments appeal to total aggregate wellbeing, average wellbeing, or wellbeing above a certain critical level, as well as around questions about whether moral assessments reflect equal versus unequal distributions of wellbeing. A signature gesture of these moral theorists is insisting that their research programme is extremely difficult, presenting participants with nearly intractably vexing problems.14 But the issues that trouble population ethicists presuppose their methodologically abstract, calculative approach to people and circumstances. Their conundrums don’t arise for moral thinkers who reject this method as unsuited to the subject matter. What distinguishes longtermism from other positions within population ethics is a pair of related claims, one empirical and the other ethical. The empirical claim is that we live ‘at a time uniquely important to humanity’s future’ in which ‘major transitions in human history have enhanced our power and enabled us to make extraordinary progress’ while also putting us at risk of self-annihilation.15 The ethical claim has to do with what population ethicists call ‘the intuition of neutrality’, that is, the intuition that what matters morally is the quality of peoples’ lives, not how many people there are. Thinkers who incline toward neutrality hold that whether a greater or smaller number of people live at a given time is in itself morally neutral. Longtermists in contrast reject this notion of neutrality, maintaining that any additional person who lives makes the world better, as long as the person enjoys adequate wellbeing. This is the ethical backdrop against which longtermists’ empirical claim about humanity standing at a historical ‘precipice’, a time both of great promise and of increased risk of auto-extinction, seems momentous. Now it appears that a circumstance in which human beings die out in a few thousand years is worse, by many orders of magnitude, than one in which trillions of humans live on to flourish in the distant future. It appears that it would be a massive moral achievement to improve the prospect of avoiding extinction by even a fraction of a percentage. The endeavour would be so important that it would justify almost any means, however seemingly callous or appalling, including steps that resulted in the near-term **suffering** and **death** of **millions**.16 Not that all longtermists explicitly contemplate extreme or violent actions to avoid existential disasters.17 Even those who actively oppose such measures, however, offer frighteningly few safeguards to keep their moral calculations from echoing the reasoning of **murderous** **dictators** and sci-fi villains. Empty ethical equations Longtermists’ turn to existential risk marks a dramatic shift from the concern with present and near-term suffering that is the hallmark of their effective altruist progenitors. Unsurprisingly, some advocates of EA are fiercely critical of longtermism. That includes Peter Singer, whose contributions to utilitarian ethics were EA’s original inspiration. Singer is skeptical about whether humanity is indeed at a uniquely portentous moment in history, and he de-emphasises existential risk in a manner that indicates impatience with longtermists’ commitment to the posture they call non-neutrality. His aim is to redirect attention back to EA’s accent on suffering now and in the short-term. ‘If we are at the hinge of history’, he writes, ‘enabling people to escape poverty and get an education is as likely to move things in the right direction as almost anything else we might do; and if we are not at that critical point, it will have been a good thing to do anyway’.18 Singer proposes to strip longtermism of the claims that differentiate it from EA, leaving a future-oriented outlook that might be described as a generic position within population ethics. Such a future-directed EA would, he suggests, be an authoritative guide to doing good for human beings to come. But this suggestion reflects a fundamentally limited diagnosis of what ails longtermism. Even without the claims that lead its advocates to wrongly represent existential risks as swamping other moral concerns, the tradition is incapable of furnishing an understanding of our social circumstances that could responsibly inform future-oriented action. The grounds for this more negative appraisal of longtermism can be found in one of the most well-known critiques of EA. Since EA’s inception, critics have noted that its emphasis has been on assessing single action types (e.g., medical, public health, or educational interventions) in terms of the sort of wellbeing grasped by the metrics of welfare economics. They have observed that EA’s slant toward welfarism is at the same time a slant away from questions of justice, and they have revisited in reference to EA a classic charge against utilitarianism. The charge’s thrust is that EA is politically corrosive because it neglects the structural roots of global misery and so weakens political bodies capable of challenging those structures, ensuring the regular reproduction of suffering.19 Some effective altruists respond to this critique by arguing that, even if EA has in practice veered toward welfarism, there is in principle nothing to keep it from evaluating social movements’ coordinated efforts to fight for more just social arrangements and also nothing to prevent it from using the kinds of qualitative metrics that we find in disciplines in the social sciences, such as sociology and political theory.20 But this rejoinder falls flat. It is undercut by effective altruists’ reliance on the god’s eye moral method that seems to enable them to quantify values across space and species and arrive at aggregative judgments of ‘most good’. This methodological stance disqualifies anyone who adopts it from discerning the systematic injustices targeted by social justice movements. When participants in anti-racist, feminist, and Indigenous rights movements protest sustained physical, psychological, and political violence against specific oppressed human groups, they are moved by structural obstacles to flourishing and the absence of reparations. It is not possible to adequately grasp the nature of such wrongs without an appreciation of the history and function of the social mechanisms that reproduce them. Attempts to understand these injustices, when approached in the abstract and aperspectival manner characteristic of EA, uninformed by pertinent historical, cultural, and political considerations, are bound to misfire. They also risk strengthening the oppressive structures in question because one way in which these structures function is by obscuring the historically and socially specific suffering of the oppressed. That, then, is why EA is unable to slough off the allegation that it has a politically conservative, welfarist bent. It lacks the immanent resources necessary for illuminating systematic injustice and envisioning appropriate remedies to it.21 Longtermism is tainted by the same lack. Population ethics, the original home of longtermism, is premised on the assumption that, appropriately specified, an abstract account of an action’s effects on the lifetime wellbeing of prospective populations equips us to answer questions about the action’s rightness. This assumption is false. Instead of making it possible to determine what counts as right action, a detached approach obscures from view just and unjust relationships that are part of these determinations’ lifeblood. Because the calculative enterprise in which population ethicists are engaged is based on false presuppositions, the technical headaches with which it presents them are at bottom self-inflicted injuries. The correct attitude to their disciplinary puzzles is to dissolve not solve them, and this applies to the debate, central for those population ethicists who selfdenominate as longtermist, about whether to affirm an ‘intuition of neutrality’. The debate’s conceit is that there is a coherent abstract question about whether creating more happy people is a moral gain. But longtermists’ assertion of non-neutrality is nothing more than an empty gesture. The emptiness extends to the non-neutrality based computations that seem to support longtermism’s insistence on regarding existential risk as a great or even overwhelming moral priority. Longtermists’ distinctive moral math simply falls apart.22 Exploiting existential angst This isn’t yet an adequate inventory of longtermism’s major weaknesses. Once we set aside the morally freefloating calculations on which longtermists build their case for an extreme prioritising of existential risks, it might seem that we retain the makings of a helpfully future-oriented practical programme. That is the gist of Singer’s proposal for exchanging longtermism’s fixation on historical precipices with a forward-looking, utilitarian-themed project that recentres longtermism’s origins in EA. But this is a non-starter. It fails to register that EA itself is incapable of shedding light on unjust and harmful social structures or assessing efforts to resist them. Even shorn of its wrongheaded stress on existential hazards, longtermism is a treacherous guide to acting responsibly towards those who will come after us. This emerges concretely in MacAskill’s and Ord’s treatments, in their respective recent books, of climate change and other forms of anthropogenic environmental destruction. Both discussions are distorted by a disturbing interest in existential risk that makes it seem per tinent to investigate whether global heating will lead to human extinction or whether it will ‘only’ kill billions of humans and trillions of animals and devastate ecosystems, while still permitting the survival and ultimate flourishing of small human groups.23 This misguided preoccupation with existential dangers is closely tied to other outrages, such as MacAskill’s selective and highly contentious appeal to climate science in support of his chillingly casual ‘best guess’ that some human beings would survive ‘fifteen degrees of warming’.24 But, even apart from their morally disastrous hang-ups with human extinction, MacAskill’s and Ord’s reflections on the environmental crisis are ruinously wrongheaded. When they consider strategies for reducing greenhouse gas emissions to combat the devastation of climate change, they limit themselves to strategies that can be pursued within existing socioeconomic arrangements. This includes technological innovations such as ‘clean’ or low carbon energy sources and different forms of geoengineering.25 It also includes policies such as internationally coordinated emission-reduction schemes.26 MacAskill at one juncture mentions youth activism admiringly, but his point about it is simply that it can increase public support for climate pledges.27 Nowhere in Ord’s or MacAskill’s remarks is there any real acknowledgment of the reality, repugnant to members of the billionaire class they assiduously and successfully cultivate, that meaningful environmental action will need to involve new values and substantial social change.28 Still more striking, perhaps, is that MacAskill and Ord try to diminish our sense of the urgency of environmental issues, arguing that we should regard renegade AI and human-developed pathogens as more critical because likelier to trigger human extinction. This line of argument, common among longtermists, is a further expression of the warping moral effects of a fascination with extinction risks, which seems to speak for downgrading the exigency of things that don’t extinguish human life altogether, and so supports treating as relatively morally insignificant the terrible fact that huge numbers of people are already dying, being uprooted from their communities, and suffering other great hardships because of climate change.29 Yet, even within the context of MacAskill’s and Ord’s extinction-focused programme, it is not clear why the environment fails to loom larger. Ord argues that environmental degradation is relatively unlikely to directly produce an extinction event and more likely to generate forms of political instability that indirectly lead to one, providing the conditions for other anthropogenic dangers.30 It’s not clear why that should make it less imperative to attend to environmental factors, or why a deviant robot takeover should be a bigger priority, unless it’s just that, considered in isolation, deviant AI appears be a hazard addressable with the kinds of instruments that Ord and other longtermists have at their disposal. Here the drive to downplay the seriousness of environmental crisis plainly outruns the grounds for doing so. Longtermism is marred not only, therefore, by a misjudged positioning in population ethics that swings it toward existential risk but also by methodological presuppositions that prevent it from recognising that movements for social change, such as the environmental movement in its interplay with anti-racist and other social justice movements, have long been engaged in the kind of future-facing social enterprise it preposterously credits itself with inaugurating.31 These objections are not at base about the troubling fact that the tradition is the brainchild of a group of white men at an elite university, some of whom have records of racist statements.32 More fateful is a dimension of longtermism’s signature theories of existential risk. These theories treat as less urgent those anthropogenic hazards that won’t snuff out humanity altogether, and the theories’ adherents place the currently intensifying human-caused climate crisis squarely in this category, encouraging us to regard as morally less important the suffering and death it is occasioning. The harms in question are falling in dramatically lopsided fashion on racialised and Indigenous groups the world over, groups whose very vulnerability to these harms is a product of long histories of injustice. Such theory-induced callousness to losses and damages visited grossly unequally on racialised people licenses talk of a racist strain in longtermist thinking, and individual longtermists **deepen** this strain in specific ways. A well-placed young longtermist once argued that inhabitants of rich countries are generally more ‘innovative’ and ‘economically productive’ and that saving their lives is hence substantially more important for humanity’s future than saving lives in poor countries.33 Today some of the tradition’s most prominent champions advocate projects of bio-enhancement, reminiscent of twentieth-century eugenics, aimed at developing a transhuman species that is better equipped for survival in the long-term.34 These sorts of reinforcements of longtermism’s racist streak are onlystrengthened by the tradition’s inability to grasp, and consequent proclivity to make invisible, contributions to revolutionary antiracist struggle.35 Mega-philanthropic delusions The story of longtermism is not just a tale of a no good, very bad moral theory. As the coffers of longtermism’s institutes and related charities have swelled, it has begun to enact its priorities, funding research on misaligned AI and anthropogenic pathogens and supporting institution-building, with research grants as well as grants to EA’s and longtermism’s institutes.36 Its arrival as a philanthropic player exposes it to concerns about having an unmerited sway on social issues. Like other wealthy private foundations, longtermist organisations are able to specify what counts as good and shape civic life without real public answerability. In the US and elsewhere, tax exemptions of well-funded private charities take from the public till huge sums that voters could otherwise have directly determined how to spend, and, apart from relatively insignificant tax obligations and reporting duties, there is little accountability. This is a money-fueled arrangement involving ‘the exercise of wealth-derived power in the public sphere with minimal democratic controls and civic obligations’.37 With its growth into a movement, longtermism has joined this undemocratic commandeering of the public realm, using its financial heft to promote its dangerous obsession with existential risk. Longtermism’s moral case for accenting such risk deflects from present suffering in a manner that simultaneously absolves harmful socioeconomic mechanisms from criticism and hastens the sorts of hazards it is supposed to head off. Yet it has been singularly successful at attracting rich backers to its project. In treat ing the economic arena to which these individuals owe their wealth as critically off limits, it positions them to look upon themselves, not as complicit in the arena’s injustices, but as singled out by their success in it to be world saviours.38 A deceitful narrative of selfless heroes riding to humanity’s rescue has proven ideologically effective, and it seems clear that many longtermists – students, researchers and members of the public, as well as donors – are sincerely committed to what they take to be a uniquely important moral enterprise. But their sincerity is no argument against the corruption of a movement that uses a bankrupt morality to justify profiting from the systems most threatening to the future it claims to secure. The fact that some major supporters of longtermism, such as Bankman-Fried, have been suspected of financial fraud is a sideshow to the main event. Longtermism’s corruption is **inseparable** from the way in which its core ideas are put into practice, and the baseness is still there when its programmes are pursued with rigorous legality. A critique of longtermism that enabled its adherents to see it in this harshly revealing light would be a welcome step towards envisioning and enacting a just and livable future.39

#### Err neg – current AGI risk literature is faulty

McLean et al. 21 [Scott McLean, researcher at the Centre For Human Factors And Sociotechnical Systems at the University Of The Sunshine Coast, Gemma J. M. Read, researcher at the Centre For Human Factors And Sociotechnical Systems at the University Of The Sunshine Coast, Jason Thompson, researcher at the Centre For Human Factors And Sociotechnical Systems at the University Of The Sunshine Coast and researcher for the Transport, Health and Urban Design (Thud) Research Lab at the University of Melbourne, Chris Baber, researcher at the School Of Computer Science at the University Of Birmingham, Neville A. Stanton, researcher at the Centre For Human Factors And Sociotechnical Systems at the University Of The Sunshine Coast, and Paul M. Salmon, researcher at the Centre For Human Factors And Sociotechnical Systems at the University Of The Sunshine Coast, 1-20-2021, "The risks associated with Artificial General Intelligence: A systematic review", Taylor & Francis, https://www.tandfonline.com/doi/full/10.1080/0952813X.2021.1964003#d1e426]/Kankee

Conclusion The current systematic review was conducted to investigate the extant peer reviewed literature focused on the risks associated with AGI. Data extracted from the eligible articles included, the type of analysis methods used, risks associated with AGI, and recommended risk controls/risk management strategies. From the small number of eligible articles, a broad range of risks were identified including, AGI removing itself from the control of human owners/managers, AGIs being given or developing unsafe goals, development of unsafe AGI, AGIs with poor ethics, morals and values, inadequate management of AGI, and existential risks. However, issues with the current state of peer reviewed AGI risk literature emerged. First, there was a scarcity of modelling techniques applied to investigate risks associated with AGI. Second, there was a limited number of studies that focused on the AGI risks in specific domains. Third, the lack of information regarding the AGI systems considered in terms of specifications, goals and tasks raises questions about the validity and comprehensiveness of the risks identified. Fourth, there was a limited amount of peer reviewed literature on the risks of AGI. Finally, there is a lack of consensus on the terminology used within AGI research. It is concluded that there is a critical need to address the multiple issues identified in the current review. Given that the fate of humanity may be at stake with the development of unsafe AGI, it is essential that we have reliable, valid, and rigorous research to guide safe AGI design, implementation and management.

### Contention 7: Veil of Ignorance

#### Rigorous application of the veil of ignorance, free from bias, produces maximally egalitarian and fair outcomes, dignifying everyone with intrinsic worth

Davies 19 [Ben Davies, research fellow at the Uehiro Centre for Practical Ethics at the University of Oxford, 2019, “John Rawls and the “Veil of Ignorance’”, Oklahoma State University Library, https://open.library.okstate.edu/introphilosophy/chapter/john-rawls-and-the-veil-of-ignorance/]/Kankee

John Rawls’s Veil of Ignorance is probably one of the most influential philosophical ideas of the 20th century. The Veil of Ignorance is a way of working out the basic institutions and structures of a just society. According to Rawls, [1], working out what justice requires demands that we think as if we are building society from the ground up, in a way that everyone who is reasonable can accept. We therefore need to imagine ourselves in a situation before any particular society exists; Rawls calls this situation the Original Position. To be clear, Rawls does not think we can actually return to this original position, or even that it ever existed. It is a purely hypothetical idea: our job in thinking about justice is to imagine that we are designing a society from scratch. The idea is that social justice will be whatever reasonable people would agree to in such a situation. We can then start thinking about how to make our actual society look more like the ideal picture we have imagined. Of course, if we were designing a society in the Original Position, people might try to ensure that it works in their favour. The process is thus vulnerable to biases, disagreements, and the potential for majority groups ganging up on minority groups. Rawls’s solution to this problem comes in two parts. Firstly, he makes some assumptions about the people designing their own society. People in the Original Position are assumed to be free and equal, and to have certain motivations: they want to do well for themselves, but they are prepared to adhere to reasonable terms of cooperation, so long as others do too. Rawls also simplifies his discussion by imagining that people in the Original Position do not have total freedom to design society as they see fit. Rather, they must choose from a menu of views taken from traditional Western philosophy on what justice involves. The second part of the solution is the Veil of Ignorance. This involves a further leap of imagination. When we are thinking about justice, Rawls suggests that we imagine that we do not know many of the facts – both about ourselves and the society we currently live in – that typically influence our thinking in biased ways. By intentionally ignoring these facts, Rawls hoped that we would be able to avoid the biases that might otherwise come into a group decision. For instance, if I were helping to design a society, I might be tempted to try to make sure that society is set up to benefit philosophers, or men, or people who love science fiction novels. But if I don’t know any of those facts about myself, I can’t be tempted. The Veil is meant to ensure that people’s concern for their personal benefit could translate into a set of arrangements that were fair for everyone, assuming that they had to stick to those choices once the Veil of Ignorance ‘lifts’, and they are given full information again. One set of facts hidden from you behind the Veil are what we might call ‘demographic’ facts. You do not know your gender, race, wealth, or facts about your personal strengths and weaknesses, such as their intelligence or physical prowess. Rawls thought these facts are morally arbitrary: individuals do not earn or deserve these features, but simply have them by luck. As such, they do not deserve any benefits or harms that come from them. By removing knowledge of the natural inequalities that give people unfair advantages, it becomes irrational to choose principles that discriminate against any particular group. The Veil also hides facts about society. You do not know anything other than general facts about human life, and in particular you do not how their society is organised. Finally, the Veil hides facts about your “view of the good”**:** your values, preferences about how your own life should go, and specific moral and political beliefs. Rawls was a political liberal. That meant, among other things, that he thought the state should be neutral between different views about value. So, Rawls isn’t afraid to make several significant assumptions about the people involved in making decisions behind the Veil. Some of his assumptions aim to turn the conflicts that arise between self-interested people into a fair decision procedure. As we’ll see, however, others might be more fairly criticised as unreasonably narrowing the possible outcomes that people can reach behind the Veil. I will outline Rawls’s justification for the Veil of Ignorance, raise some potential challenges for the conclusions he thinks people will reach from behind it, and lastly consider three criticisms of the Veil of Ignorance as a theoretical device. While these criticisms differ in their substance, they are united by a common feature: their scepticism of the way the Veil abstracts from real life in order to reach conclusions about justice. I’ll conclude that these criticisms have merit; the Veil of Ignorance, considered by itself, does lead us to ignore the real world too much. However, I’ll suggest that, at least in their strongest versions, these criticisms miss an important benefit of the Veil: quite simply, the fact that our own personal concerns and values can bias our thinking about justice, and that we can make important progress by considering things from different points of view. The principles of justice Imagine that you find yourself behind the Veil of Ignorance. You might want to make sure that your life will go well. If you had to design a good life for yourself, you’d go for the specific things you care about. But behind the Veil you don’t know those specifics; you only know things that generally make people’s lives go well. Rawls calls these ‘Primary Goods’. They include things like money and other resources; basic rights and freedoms; and finally, the “social bases of self-respect”: the things you need to feel like an equal member of society. In Rawls’s view, a central challenge behind the Veil is the lack of probabilities available. If you knew that your society was 90% Catholic, you could set things up so that the rewards associated with being Catholic were much higher. That would be personally rational, since you are very likely to end up in the better off group. The Veil prevents this type of reasoning because it hides the information. In the complete absence of probabilities, Rawls thinks you should play it safe and maximise the minimum you could get (a policy he calls Maximin). Translated into a society, that means that we should ensure that the worst-off people in society do as well as possible. Rawls suggests two principles will emerge from discussion behind the Veil: First Principle: Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, compatible with the same liberties for all; Second Principle: Social and economic inequalities must be: Attached to offices and positions open to all under fair equality of opportunity; To the greatest benefit of the least-advantaged members of society (the difference principle). Rawls opts for equality of basic liberties in the First Principle because he thinks this is essential for seeing yourself as a moral equal in society. For other Primary Goods, though, equality is less important. By allowing some inequality, we could make life better for everyone. If we attach higher salaries to certain jobs, they may attract the hardest working people, producing greater economic benefits for everyone. The two parts of Rawls’s second principle of justice set limits on when inequalities are allowed. Fair equality of opportunity says that positions which bring unequal payoffs must be open to people of equal talents and equal willingness to use them on an equal basis. If two people are just as capable of doing a job, and just as hardworking and willing to apply themselves, neither should have a greater chance of securing the position because they are wealthier, or because of their race or religion. Of course, we might wonder (and Rawls does not give a clear answer about this) when we are supposed to judge whether two people are equally hardworking and talented. The talents you choose to develop, and the amount of effort you put in, are heavily affected by education; so it might seem unfair to judge people if they have had very different educational experiences. Rawls’s argument therefore seems to support ensuring broad equality of education, encouraging people to find and develop their talents to the fullest, even if this isn’t a conclusion he explicitly draws. Finally, the Difference Principle sets a further restriction on inequalities. Even if a particular inequality does not affect equality of opportunities, the Difference Principle tells us that it must be beneficial for the very worst off. For instance, it might be that by allowing inequalities, we motivate people to work harder, generating more Primary Goods overall. If these then benefit the worst off in society, making them better off than they would have been in a more equal distribution, the Difference Principle will allow that inequality. Criticisms

#### Bias is inevitable in humans and harms principles of universalizable decision-making – mitigation fails

Jones 19 [Pete Jones, chartered Psychologist and Chartered Scientists in the UK, 01-25-2019 “Unconscious Bias – avoidable or inevitable?”, European Research Council, https://erc.europa.eu/news-events/magazine/unconscious-bias-–-avoidable-or-inevitable]/Kankee

Can you give us an example of implicit bias? The most prevalent effect of unconscious bias is affinity bias. We have an inclination to prefer people who are similar to us on the basis of a wide range of characteristics, including social or career background, gender, accent, education, ethnicity, age, hobbies and interests, etc. But there are many many other biases. Anchoring can result in relying too heavily on one piece of information or a personality trait when making decisions. For example, if an applicant has a degree from a particularly high quality institution, the qualification becomes the entire or main focus of the employment decision. Stereotyping can create judgements about individuals within particular social groups. Confirmation bias can mean that once we feel someone or a situation is a particular way, we seek out information to confirm it and ignore evidence to the contrary. Research and measurement of bias To understand how and when to try and avoid a particular bias we first need to establish whether it actually exists and also how strong it is. In 1868 the Dutch physician Franciscus Donders tried to measure differences in human reaction time to infer differences in cognitive processing. Ultimately this could be used to measure bias by examining the speed with which bias was shown in the associated process of decision making – e.g. how quickly a subject would associate the words "man" with the word "scientist" and so exhibit gender bias. Technological limits of the time made it impossible to measure the milliseconds involved. In 1950s America unconscious bias was often discussed in terms of racial prejudice. Social psychologists such as Gordon Allport worked to define implicit bias, and work carried out in 1998 at Harvard created a popular way to measure the success or failure of action against it: the Implicit Association Test (IAT). This process became established as a metric of choice in clinical forensic psychology but it is a still subject to debate in academic circles due to inconsistent results. Benchmarking data from previous studies is another way to estimate bias. One study implies that 30% of academic staff has sufficient bias in the areas of ethnicity or disability to influence their behaviour. Can we eliminate bias? No, but a realistic aim is to try and find ways to mitigate bias and so avoid distortion in decision making. Research from Harvard found that the effects of personal interventions such as awareness raising at a personal level are positive, but short-lived. Giving people strategies to use to mitigate bias proved more effective, resulting in a downward trend in bias displayed over a three month period. It has been estimated that even one hour of awareness training, plus understanding and putting mitigation strategies into practice, has a positive effect. Systems used for decision making should also be adapted to minimise the chance that bias can influence the process. But ultimately we need to take a certain amount of personal responsibility. As with many things in life, it's down to us how we behave in practice.

#### AGIs are comparatively better decision-makers, fully able to utilize the Veil of ignorance to equitably make unbiased decisions

Olsher 24 [Daniel Olsher, researcher and the founder of Integral Mind Technologies LLC, 01-30-2024 “Proof of Achievement of the First Artificial General Intelligence (AGI)”, HAL Open Science, https://hal.science/hal-04397466/document]/Kankee

Provably Unbiased In order to deliver all of the foregoing, any true AGI must be fundamentally and provably unbiased. It must be possible to show ahead of time that bias cannot and will not be present, as a biased system would not be useful and could not be safe nor trustworthy. Not only would it be impossible to predict where and how biases might manifest, but clear and unbiased understanding is required in order for systems to be able to think properly and conduct accurate simulations. Biases also necessarily prevent systems from properly delivering intelligent responses in certain situations, thus rendering them unable to meet core AGI criteria. As we show below, the system presented here has identified and removed each potential source of bias in advance and its processes and procedures prevent bias from ever entering the system in the first place. At a minimum, because It is not statistical, does not use models, and does not employ training data, all biases resulting from these sources, including implicit human biases, are removed. Independent Consciousness In order to truly be free of bias, any genuine AGI must not rely on human cognition. Humans experience over 200 different categories of bias for which no remediation is possible. In addition, as we explore below, any level of human dependency renders AGI systems unusable in the real world and prevents them from being able to compete with other superintelligent systems in practice. Reality is fundamentally too complex for humans to comprehend. In response, we evolved consciousness, which focuses limited cognitive resources on just those elements of situations that appear to be most relevant to survival while ignoring everything else. But this process creates profoundly false understandings of reality. What actually matters most in any given situation changes from moment to moment and always depends on context, but consciousness never quite catches up. Consciousness causes us to see the world in arbitrary ways. It misunderstands and misconstrues much of reality and often suffers from confusion caused by viewing the world through stovepipes instead of as a whole. Assumptions build up, and while these are often incorrect, they are useful enough in everyday life, and are shared by enough people, that they tend to persist. But when making decisions, and when generating safely intelligent behavior, these assumptions damage every element of decisionmaking. They turn arbitrary accidents of history, process and personnel into key determinants of decision quality. Humans aren’t aware of their biases and are unable to determine when their hidden assumptions and beliefs no longer apply. They often fall into ‘tunnel vision’. As a consequence, traditional decisionmaking processes ignore far too much key information and are unable to adjust at the rate changing circumstances require. It becomes impossible to accurately see, predict, or understand, a consequence of which is that outcomes are dictated by chance, not choice. Any genuine AGImust be able to see the world clearly and as it truly is, without dependence on humans and without bias, as only in this way will it be able to understand exactly what is happening (and why) and in so doing create order from chaos. Biases sharply reduce the amount of information that can be accurately taken into account and introduce incorrect information, but this is entirely unacceptable if we wish to achieve the real‐world success intelligence, and thus AGI, requires. In addition, as noted above, the mere existence of AGI leads to exponential leaps. We all hold unconscious mental models that tell us what to expect from the world, but our models in fact implicitly depend on the hidden assumptions that current levels of intelligence, speed, and human competency will continue to hold in future. Superintelligence entirely shatters all such assumptions. Those who possess superintelligence are able to think, understand, decide, and act faster, better, and much more powerfully than everyone else, meaning that the human limitations implicit in current models no longer hold. The floor raises, and those who continue in old thinking are no longer able to compete. Once anyone has superintelligence, all must have it simply in order to maintain the status quo. Everything is therefore different from that point on. The effects of superintelligence are especially confounding because they are strongly non‐linear and take place across multiple stovepipes, meaning that no one person or group can easily see or understand them. Small changes in one part of the system exert major, unobvious, effects in apparently unrelated areas. People tend to hold on to their beliefs far past the time they have ceased to be useful, but this is no longer tenable in an AGI world. Anyone choosing to ignore new realities will be unable to win due to the immense advantages obtained by using better systems.