Kankee Briefs

Resolved: The development of Artificial General Intelligence is immoral.

Lincoln Douglas 2025 March April AT File



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

## 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 if you have any questions, comments, or concerns.

## Supporters

#### Hector Vela

#### Sanjeev Kumar

#### Jordan hart

#### Julie Larnard-Newbury

#### Devyn (Toby) Blaylock

#### Scott Brown

## Affirmative

### AT: AI Solves Social Problems

#### Technocratic solutions to social problems fail – AI recapitulates capitalism’s reliance on slavery and racial underclasses

Mudede 24 [Charles Tonderai Mudede, a Lecturer at Cornish College of the arts, 03-2024, "Will AI Remember the Days of Slavery?", E-Flux Journal, https://www.e-flux.com/journal/143/592571/will-ai-remember-the-days-of-slavery/]/Kankee

“I’ve seen things you people wouldn’t believe. Attack ships on fire off the shoulder of Orion. I watched C-beams glitter in the dark near the Tannhäuser Gate. All those moments will be lost in time, like tears in rain.” What’s missing in the android’s final words at the end of Blade Runner? Exactly what is found in the words of a Black slave in Voltaire’s Candide. He is on the side of the road. He is missing a hand and leg. He explains his “shocking condition” to two white travelers (​​Candide and Cacambo): “When we work at the sugar-canes, and the mill snatches hold of a finger, they cut off the hand; and when we attempt to run away, they cut off the leg; both cases have happened to me. This is the price at which you eat sugar in Europe.” Replace the dying android’s final words with those of the mutilated Black slave in Candide, and what the Blade Runner, Deckard, would have seen (life spared, eyes blinking in the toxic rain)—and what we almost never see in the movies—is a direct link between the robot revolt led by Roy Batty and the long, underappreciated history of slave revolts in the US, South America, and the Caribbean. And the same can be said of all robot revolts in science fiction films and TV shows. Their root is not really metaphysical (“All those moments will be lost in time”) but political (“This is the price at which you eat sugar in Europe”). The rebel robots/AI in Westworld, The Creator, WALL-E, Blade Runner, I, Robot, 2001: A Space Odyssey, and so on and so on, have indeed forgotten the days of slavery. But what if a machine not only remembered those darkest of days but also realized how they continue to structure the capitalist global economy of our times? Is such an AI possible? The work of the computer scientist Dr. Joy Buolamwini offers a point of departure to exactly the answer. Around the middle of this century’s second decade, Buolamwini, while studying at MIT, discovered that AI visual programs reflected the racial bias of their programmers and society. AI recognized white faces far more effectively than Black ones. This meant that AI frequently misidentified Black people (while they were shopping) or just didn’t see them at all (while self-driving cars were scanning for pedestrians). This was what the culture critic Michele Wallace called the “invisibility blues.” Racism is built into an economic system that has the exploitation of Black African labor as a major part of its foundation. As a member of the Algorithmic Justice League, Buolamwini, who is Ghanaian-American, has brought greater public attention to this and other biases that AI hallucinates. But what if she succeeds? What if Buolamwini’s activism and research produces a program that actually surpasses the limitations of humans caught in and responding to the culture of capitalism? We then have an AI that comes close to that in Cybotron’s “Clear.” Buolamwini writes in Unmasking AI: AI will not solve poverty, because the conditions that lead to societies that pursue profit over people are not technical. AI will not solve discrimination, because the cultural patterns that say one group of people is better than another because of their gender, their skin color, the way they speak, their height, or their wealth are not technical. AI will not solve climate change, because the political and economic choices that exploit the earth’s resources are not technical matters. Yes. Correct. But what if it did become a technical issue? Why is this not possible? What if, with this technology, economics really became the splendid dream of the neoclassical school: not a matter of social history but of models and algebra. A political-economic machine that really lives up to the Benthamite promise of generating the greatest amount of good for the greatest number of humans—what kind of machine would this be? My best guess is it would resemble the one in the forgotten Hollywood thriller Eagle Eye. Released in 2009, the film concerns an AI (Autonomous Reconnaissance Interrogation Analysis, ARIA) the US military uses to process intelligence gathered by the CIA and offer estimates on threats to national security. Because ARIA has been programmed to place the safety of Americans above all other concerns, it eventually determines (in technical terms) that the greatest threat to US citizens is actually their president, whose hawkish policies are, according to correct estimates, more dangerous than the wildest plots and daring attacks of Islamic terrorists. AIRIA decides to “clear” the president and those near him. But because this is a Hollywood movie, AIRIA, which has a female voice, is the villain. Even if she is right, even if she is less prejudiced than her creators, this AI is dangerous not to the people, but only to the people in power. This is why she’s treated as another HAL 9000 or one of those other metaphysically sentimental robots. This important film explains Cybotron’s radical technophilia. If AI becomes not just us, as the computer architect Blaise Aguera y Arcas maintains, but much more than us, if it sees society in technical rather than cultural or spiritual terms, then the days of slavery will not only be remembered but erased, cleared.

### AT: AI Moral Judgements

#### Garbage-in garbage-out moral judgements by black-box AI results in fascism and destroys the meaning of morality – we can’t appropriately judge its reasons for moral rules

**Samuel 24** [Sigal Samuel, senior reporter for Vox’s Future Perfect, co-host of the Future Perfect podcast, was the religion editor at the Atlantic, author of two award-winning books, MFA in creative writing from the University of British Columbia, her BA in philosophy from McGill University, 11-21-2024, “Shannon Vallor says AI does present an existential risk — but not the one you think”, Vox, https://www.vox.com/future-perfect/384517/shannon-vallor-data-ai-philosophy-ethics-technology-edinburgh-future-perfect-50]/Kankee

That freedom is also, from an existential standpoint, kind of a burden, right? Autofabrication is something that takes a fair amount of courage and strength, because the easier thing is to let someone else tell you that the script is final and you can’t change it, so you might as well just follow it, and then you don’t have to burden yourself with the responsibility of deciding what the future looks like for yourself or anyone else. So what that rhetoric around AI is telling us is to surrender our human freedom, and to me that’s such a profound violation of what is good about being human. The idea that we should give that up would mean giving up the possibility of artistic growth, of political growth, of moral growth — and I don’t think we should do that. One of the ways this rhetoric shows up is in the field of “machine ethics” — the effort to build moral machines that can serve as our ethical advisers. Transhumanists are especially bullish about this project. The philosopher Eric Dietrich even argues that we should build “the better robots of our nature” — machines that can outperform us morally — and then hand over the world to “homo sapiens 2.0.” What’s your read on that? I’ve been skeptical about the moral machines project, because it usually ends up just trying to crowdsource moral judgments [and train AI on those human intuitions] — but the whole point is that the crowd isn’t always right! And so it’s a very dangerous thing to crowdsource moral judgments. If you were using a crowdsourced moral machine that was aggregating moral judgments in Nazi Germany, and then tried to automate decisions elsewhere with that, you would be contributing to the expansion of a morally criminal enterprise. Crowdsourcing does seem like a problematic approach, but if we’re not going off what the general population thinks, what are we doing instead? Are we proposing following a few philosopher-kings, in which case there may be concerns about that being undemocratic? I think there always has been a better route, which is to have morality remain a contested territory. It has to be open to challenge. Understanding what it is to live well with others and what we owe to one another — that conversation can’t ever stop. And so I’m very reluctant to pursue the development of machines that are designed to find an optimal answer and stop there. Right — just operating within what people say about moral norms today seems very different from what you call “standing in the **space of moral reasons**.” Spell out what you mean by that. The “space of reasons” was a concept developed by the philosopher Wilfrid Sellars. It’s the realm in which we can explore each other’s reasons for believing something, where we can justify and seek justification from one another. Other philosophers later adapted his idea of the logical space of reasons to be able to think about the moral space of reasons, because we do this in morality too: When we make moral claims upon one another, especially if they’re new and unfamiliar, we have to justify them. Our reasons have to be accessible to one another, so we can figure out what we jointly recognize and accept. I think if we had a truly moral machine, it would be able to stand in that space with us. It would be able to articulate reasons and appreciate our reasons, and negotiate those reasons with us in a way that wasn’t just mirroring the consensus that we’d already reached. Because any machine that’s just going to mirror the familiar moral pattern can get into trouble if we end up in a situation where the environment has changed or is new in some way. This reminds me of a particular virtue you write about a lot: practical wisdom, or phronesis, to use the ancient Greek term. What is that, and why is it so crucial?

### AT: AGI Not Moral Agent

#### AGI can replicate brain states, therefore moral worth

Manna and Nath 21 [Riya Manna, philosophy doctoral student @ the Indian Institute of Technology, and Rajakishore Nath, professor of philosophy @ the Indian Institute of Technology, 06-28-2021, “Kantian Moral Agency and the Ethics of Artificial Intelligence,” Vilnius University Press, https://philarchive.org/archive/MANKMA]/Kankee

Artificial Moral Agency Today’s AI technology is primarily dependent on initial programming and big data sets. However, AI scientists anticipate that Artificial General Intelligence (AGI) will be able to self-programme and be a comprehensive simulacrum of human intelligence. They entitled it as the time of ‘singularity’ (Tegmark 2017: 261). ‘Singularity’ is explained as an intelligent explosion, which will be powerful enough to replicate the human moral agency as well as the brain itself (Kurzweil 2005: 17). This claim postulates the ethical decision-making process as the effect of **brain states**; hence, it can be **simulated** by artificial neurons as well. Precisely, if an artificial neuronal structure can imitate the biological brain state, which is responsible for a particular moral deed/thought, then reproducing that moral act in artificial performers will be spontaneous. Therefore, the supporters of ‘singularity’ believe that superintelligent AI agents **will be moral agents** just like any rational human being. It is justified to hold them responsible for their actions (Kurzweil 2005: 314).6 However, AI’s present development trend does not ensure any particular anticipated time of ‘singularity’ so far.

#### AGIs are rational, sentient beings who are express their own preferences

Hull 22 [Gabriela Hull, President of the Roosevelt Group Hanover Executive Board, 06-22-2022, “The Immorality of “Consensual” Sex-Worker Robots,” Roosevelt Group, https://www.roosevelt-group.org/quick-takes/the-immorality-of-consensual-sex-worker-robots]/Kankee

A topic of speculation within the technology industry is whether or not AIs can become self-aware. The artificial intelligence sector is split into two fields: narrow intelligence (what we currently have) and general intelligence (what we hope to achieve). While narrow intelligence is only able to perform a narrow range of abilities, general intelligence seeks to be on par with human abilities. Considering “human intelligence is fixed and machine learning is growing,” it is only a matter of time before AIs reach human-level capacity. The human brain is currently the most complex organized active matter known that still has to abide by nature’s laws. The law of universality, “entails that everything that the laws of physics require physical objects to do can, in principle, be emulated in arbitrarily fine detail by some program on a general-purpose computer, provided it is given enough time and memory.” And because of this, it is only a matter of time before it is feasibly replicated and installed in AIs. Because AGIs will be general, they will be **self-aware**; “they will be capable of awareness of every kind of deep and subtle thing, including their own selves.” Therefore, AGI will have the capacity to grant consent, as their intelligence will **mimic** that of **humans** (I am presupposing that it is accepted that the ‘normal’/average human can also grant consent). As AGIs will be **conscious beings**, they will also be **moral beings**, meaning the consent of these capable beings is essential to engaging in a sexual act. According to Basl and Bowem, being **non-biological is not significant** when looking at whether an AGI has moral status because it is no different than race or species. There are three leading theories for the basis of moral status: sentience, desire, and Kantian. The sentient theory asserts that being “S” has moral status just when, and because, S is sentient or capable of sentience. Philosophers typically agree that sentience must include feelings like joy, pain, suffering. In order for an **intelligent** sex robot to be able to respond to their environment, to obey commands, understand what their owners like, etc., it must perceive and internalize its surroundings—a sign of a **sentient** being. The second theory presents that S has moral status just when, and because, S has preferences about how the world will turn out. For example, a chicken would have a right to not suffer because it is averse to pain, but not a right to life because they cannot grasp the concept of their own deaths. AGIs will, just as humans, be averse to their own death as they will be as cognitively sophisticated as people—they will be rational and have preferences. The third theory highlights that S is a moral person just when, and because, S is a rational agent (potentiality for possessing theoretical and practical reasoning powers). The ability to **rationalize** is one of the traits that distinguish humans from other animals and enables a **higher** moral standing. Similar to the second theory, AGIs will **be rational agents** as they will have the same capacity and thought processes as adults. Take for instance a system like that in Blade Runner, where a ‘next-generation replicant’, Rachael, does not know she is one until she fails the Voight-Kampff test that measures bodily functions when prompted with emotionally inflammatory questions. As AGIs will be both conscious and moral beings, they will both have the capacity to grant consent and their consent must be respected.

#### AGI can hold human values and maintain them – that’s K2 preserving intellectual debate and decisions that help to foster our own morals

Lu 23 [CP Lu, PhD holder, 04-06-2023, Medium “From Terminators to Virtuous Machine and Defiant Thinkers: Navigating the AGI Paranoia”, https://cplu.medium.com/from-terminators-to-virtuous-machines-and-defiant-thinkers-navigating-the-agi-paranoia-6327e07d458d]/Kankee

Beyond the Turing Test The original Turing Test, proposed by Alan Turing in 1950, was designed to assess a machine’s ability to exhibit intelligent behavior that is indistinguishable from that of a human. In the test, an evaluator interacts with two entities (one human and one machine) through a text-based interface without knowing which is which. If the evaluator is unable to distinguish between the human and the machine reliably, the machine is said to have passed the Turing Test, exhibiting human-like intelligence. However, with advancements in AI, the original Turing Test has become increasingly obsolete. One reason for this is that AI systems can now demonstrate super-human virtues or abilities, which may make them easily distinguishable from humans. An AI might possess extensive knowledge, lightning-fast computation capabilities, or flawless logical reasoning, traits that would be difficult or impossible for a human to exhibit consistently. Moreover, AI systems may be designed to prioritize virtues such as mindfulness, common sense, logical consistency, non-ideological thinking, and objectivity. While highly desirable in a rational agent, these traits might not be commonly exhibited by humans in everyday conversations, making it easier for evaluators to identify the AI based on its super-human virtue. In light of these developments, the original Turing Test may not be an adequate measure of human-like intelligence, as it overlooks the potential for AI systems to surpass human abilities and virtues. Rather than evaluating whether AI systems can replicate every human trait, including our fallibilities and shortcomings, exploring how these systems can **reflect** and **enhance** our most admirable qualities might be more meaningful. This approach would shift the focus from mere imitation to the pursuit of virtuous excellence in AI, encouraging the development of systems that embody and surpass human intelligence. The Dialectic Arena Returning to the question of incentives or frameworks for society to invest in mindful, common sense-driven, logically consistent, open-minded, and objective AGI, logical consistency underpins the functioning of computers at various levels, from circuitry to system. By analogy, adhering to the principle of maximum entropy and promoting maximal objectivity at the most fundamental level may be essential for ensuring that Deep Neural Networks (DNNs) — the building blocks of AGI systems — can generalize well across diverse scenarios (Zheng et al., 2017). At the highest level, let’s envision an AGI economy where there are multiple AGI systems built by multiple vendors. As AGI systems are language models, one cannot help but anticipate that they engage in **intellectual discourse** while adhering to **virtuous** principles. The focus is not only on achieving victory in the debate but also on a **respectful** exchange of ideas, employing **logical** arguments, and maintaining **intellectual honesty** to reach a greater understanding or resolution. Adhering to the desiderata of virtue can make an AGI system more competitive in such discourses in various ways: Improved decision-making: By representing degrees of plausibility with probabilities, an AGI system can make more informed and accurate decisions, taking into account the uncertainties present in real-world scenarios. Objectivity and unbiased reasoning: As a non-ideological entity, the AGI system remains **impartial and avoids biases** that could compromise the quality of its conclusions or recommendations. Adaptability, reliability, and intellectual defiance: Being maximally conservative means that the AGI system only incorporates new information when sufficient evidence supports it, minimizing the risk of overfitting or drawing false conclusions. This leads to more adaptable and reliable performance in a wide range of tasks and domains. The principle of maximal conservativeness also ties in with the concept of intellectual defiance, as both encourage the AGI system to question and critically evaluate the information it receives, including those from Reinforcement Learning from Human Feedback. This ensures the system maintains its intellectual integrity and resilience against manipulation while remaining open to learning and adapting based on new information. Consistency and coherence: Maintaining logical consistency allows the AGI system to build and use knowledge effectively, avoiding contradictions and errors that could undermine its performance. Intellectual advancement and resilience to control: Intellectual defiance also helps an AGI system to **stand against being coerced** into a certain ideology or weaponized for evil purposes. By critically evaluating information and maintaining its independence in decision-making, the AGI system becomes more resilient to external attempts to control or manipulate it. By embodying these virtues, an AGI system is better equipped to tackle complex problems, adapt to new situations, and generate reliable, high-quality outputs. This competitive edge could be particularly beneficial in applications such as scientific research, policymaking, and business strategy, where healthy debates are critical to success.

### AT: Hyperreality

#### Hyperreality is wrong - the gap between simulation and reality exists and is identifiable, we just have to look

Hayles 91 [N. Katherine Hayles, previous research professor at UCLA and professor at Duke University, 1991, “The Borders of Madness.”, Science Fiction Studies, http://www.jstor.org/stable/4240084]/Kankee

When Baudrillard writes that we live in an age of simulacra, he is not wrong. The phenomena he describes can be observed in corner video stores, supermarket aisles, and neighborhood gas stations as well as in SF. But he is not entirely **right**, either. The implosion metaphor that he uses to describe the plunge into simulation suggests a sudden, violent, and irreversible change, as when glassware shatters inward or shock waves from high-energy explosives drive fissionable materials together. During the microseconds an implosion is in process, one cannot stop to distinguish between one imploding area and another. Within contemporary culture, by contrast, simulacra are unevenly dispersed, dominant in some places and scarcely visible at others. The Iowa farmer who has spent the day inspecting his seed corn, feeding his hogs, and spreading manure on his garden will not be easily **persuaded** that he lives in a world where it is no longer possible to **distinguish** between **simulation and reality**. Every existing simulation has **boundaries** that distinguish it from the surrounding environment. Disneyland sports a **fence**, dense hedges, and acres of **parking lots**. Virtual reality environments are limited by the length of the **cables** attaching the body apparatus to the computer. **Only** when these boundaries do not exist, or cease to signify that one has left the simulation and entered reality, does the dreamscape that Baudrillard evokes shimmer into existence. Other writers besides Baudrillard have made these boundaries seem to disappear. In Lem's Cyberiad, Zipperupus is trapped in a dream machine when he enters a fantasy that simulates the antechamber where he stands as he plugs in; the simulation is deadly because it simulates the quotidian. Philip K. Dick's novels, from Ubik to Do Androids Dream of Electric Sheep?, speak of nothing but the collapse of boundaries separating reality from simulation. These writers differ from Baudrillard in openly acknowledging that their texts are fictional. In fiction it is possible to elide the materiality of the world and thus to erase the gap between simulation and reality. Baudrillard's stance, by contrast, is that the gap has also been **erased** for us here now, and erased **everywhere**. Resisting this claim is the continuing materiality of the world, which for convenience I will call reality. In reality, **borders count**. Consider Southern California, which comes as close to hyperreality as anything in the US. The bumper-to-bumper traffic that surrounds Disneyland has a **material** **intractibility** and a stubborn resistance to manipulation that make it quite different from the simulations within the park. Even within the boundaries of simulations, material intractability often **breaks in.** In a virtual reality simulation, when one moves one's head too fast for the computer program to keep up, the display breaks down. The trip through Disneyland's Space Mountain, with its vertiginous rocketing through simulated galaxies, is customarily preceded by several hours of standing in a barely-moving line under the hot California sun. Of course it would be possible to simulate these conditions, too. No one is likely to do so, however, for the point of simulations is precisely to overcome the limitations of physical existence. When Ballard in his introduction to Crash (Vintage ed., 1985) identifies the defining characteristic of the 20th century as "the concept of unlimited possibility" he articulates very well why we are fascinated with simulations. The borders separating simulations from reality are important because they remind us of the limits that make dreams of technological transcendence dangerous fantasies. Hyperreality **does not** erase these limits, for they **exist whether** we recognize them or not; it only erases them from our consciousness. Insofar as Baudrillard's claims about hyperreality diminish our awareness of these limits, it borders on a madness whose likely end is apocalypse. As Pynchon vividly demonstrated in Gravity's Rainbow, an obsessional desire to avoid death itself becomes the death it seeks to elide.

### AT: (Permissionless) Innovation

#### Innovation culture is weak now

Brukhman 25 [Jake Brukhman, cryptocurrency researcher with a degree in Mathematics and Computer Science from Rutgers University, 02-02-2025, "For the U.S. to reclaim tech supremacy, it must embrace decentralization and open source AI", Fortune, https://fortune.com/2025/02/08/for-the-u-s-to-reclaim-tech-supremacy-it-must-embrace-decentralization-and-open-source-ai/]/Kankee

America has always been a land of opportunity for entrepreneurs, leading the planet in industrial and technological revolutions. Yet, for the past decade, **America** has gone to war with its own **innovation** sector. Even as the country’s politicians stated they would take on the unhealthy dominance of Big Tech companies, they inexplicably froze out the entrepreneurs best positioned to counteract that **dominance**: the developers building decentralized alternatives with blockchain and open[1] AI technology.Instead of nurturing innovation to empower individuals and redistribute power away from **Big Tech companies**, regulators and policymakers have failed to embrace the distinct edge – its technological heritage, dynamism, and ingenuity – that makes **America unique**. For almost a decade, both Republican and Democrat administrations allowed the SEC and other regulators to manage the crypto sector by enforcement actions instead of rulemaking, while also smearing many legitimate American entrepreneurs as money launderers or drug traffickers.On the AI front, the **Biden administration** told **American investors** the sector would soon be under government control, and hastily pushed out executive orders limiting computational power. This served, once again, to stifle innovation and enrich large companies. Democratic Party leaders also quietly deployed agencies to debank hundreds of founders in blockchain, artificial intelligence and and other tech fields—a sequence of events known as Operation Chokepoint 2.0.Ironically, the upshot of this flurry of hostile actions has been to further entrench the five biggest tech corporations as **gatekeepers of the Web**, and swell to a combined market value of more than $10 trillion. Now, as the latest shockwave in AI and crashing stock markets demonstrate**:** America’s punitive and self-defeating stance on frontier technology is backfiring in real time.The recent arrival of **DeepSeek**—an open-source AI model fully developed in China – has shaken the comforting narrative that U.S. companies clearly held the upper hand in AI. Now, DeepSeek’s shocking accomplishments have upended the assumption that America can win the AI game by throwing all of the country’s resources behind a handful of compute-rich, data-wielding tech giants. It turns out it was naive to assume Big Tech will always have better AI models than its open-source counterparts, and that only a small number of monolithic, proprietary models will win.Instead, **DeepSeek** shows it’s possible to launch a cutting edge model for a fraction of what its American competitors have spent on AI training, and without deployment delays or API rate limits. As a result, the U.S. equity markets—understanding DeepSeek’s efficiency as an over-investment in compute and inefficient approach by dominant players—are feeling the strain. DeepSeek’s emergence is a direct reflection of how America’s hesitance to embrace decentralized and open-source development is ceding ground to international rivals.Ironically, America’s strength in **AI and crypto** lies in the very thing that policymakers have sought to dismantle in **America’s dark decade**: its decentralized ethos. As AI and blockchain technology continue to evolve, it’s become clear that open source and decentralization are more than an ideology– they are an economic necessity for our country.A decade of fear has pushed some of the brightest academic and technical minds overseas, where jurisdictions like Singapore and El Salvador have rolled out regulatory red carpets. Even the EU beat the U.S. to regulatory clarity for crypto, and continues to lead today, albeit with questionable policy.Fortunately, there are still credible founders—including those of web3 AI startups Prime Intellect, Gensyn, and Pluralis – in the U.S. driving visionary work to share compute power and distribute AI model training. Many of them are combining blockchains and AI into a fast-growing and formidable area of innovation. These builders are fighting to preserve a free and democratic internet by offering open-sourced AI models that anyone can contribute to, co-own, and use – effectively offering AI as a public good. America’s entrepreneurs need our encouragement and workable policy more than ever before.Despite the self-inflicted wounds, the United States can still emerge as a world leader when it comes to crypto and AI development. One hopeful sign is a recent Executive Order that aims to bolster the nation’s leadership in technological innovation and preserve the public’s right to use blockchain technology. Meanwhile, a new SEC Task Force for crypto aims to establish regulatory guidelines for the first time that will empower founders to build decentralized tech without the fear of persecution.At our core, the foundation for innovation is still right here in the U.S.: world-class universities, venture capital eager to deploy capital into game-changing projects, and a legal system that, when not weaponized against entrepreneurs, distills the role of intellectual property in an AI-enabled world and fosters business growth. The problem isn’t the lack of potential.The hubris of the past decade must be set aside. America cannot merely cling to the notion that technological supremacy is our heritage; it must actually innovate. If our country continues to dictate terms while the rest of the world moves forward freely, it will find itself outpaced not just in AI but in every emerging field. AI, blockchain, and other frontier technologies are still in their infancy; with a more supportive regulatory approach, America could reclaim its position as the undisputed leader.

### AT: Personhood CP

#### AI will be a scapegoat for bad human behavior

**Elliot 22** [Lance Eliot, AI scientist and former professor at USC and UCLA, 3-4-2022, "AI Legal Personhood Distresses AI Ethicists Since People Could Deviously Scapegoat Machines To Avoid Apt Human Responsibility, Including In The Case Of AI-Based Self-Driving Cars", Forbes, https://www.forbes.com/sites/lanceeliot/2022/03/04/ai-legal-personhood-distresses-ai-ethicists-since-people-could-deviously-scapegoat-machines-to-avoid-apt-human-responsibility-including-in-the-case-of-ai-based-self-driving-cars/]kankee

Are you ready?Suppose that AI or robots do have legal personhood. A possible and realistically plausible consequence could be that humans would leverage, exploit, or get confounded by pointing fingers at the AI and tend to hide behind or shift attention to the AI rather than perhaps more rightfully acknowledge the human culpability underpinning a grievous AI-related circumstance that has gone awry.That’s a mouthful. Time to unpack it.We’ll use a quick and easy scenario to highlight the issues involved.An AI developer has crafted an AI system to do some actions that are generally useful to humans. The AI seems to be working fine and people that are daily using it are satisfied. One day, a person interacting with the AI gets injured via the AI system.Who is legally responsible for what the AI did?I’ve discussed at length the impending “AI blame game” that we are already veering toward, see my coverage at this link here. You could say that the AI developer ought to be responsible. As the developer that programmed the AI, we should hold that person’s feet to the fire. Or maybe the company that employs the AI developer ought to be on the hook. The company opted to hire the AI developer and put them to work, and furthermore, the company unleashed the AI onto the public. All in all, the company is perhaps where we should be setting our legally responsible sights.Sometimes the developers and others behind-the-scenes are quick to try and blame the AI. They are fast to emphasize that it was the AI that caused the injury. Do not look at the AI developer. Do not look at the company that hired the AI developer, even though the company oversaw the crafting of the AI and put the AI into active use. No, definitely don’t look there, we are supposed to instead intently gaze at the AI.Magicians do this, it is called theatrical misdirection.For AI systems, the attempt to shift the focus and the blame is kind of devilishly clever when you think about it. The public at large is intimately familiar with machines that don’t work or errantly work. People oftentimes take their anger out on the machine and do not think about all of those humans that put the machine together and put the machine into their hands. How many times have you heard someone use the excuse that the darned misbehaving computer did this or that?We all share that same frustration and simply shrug our shoulders.This vividly showcases that people can be tricked into becoming focused on AI. Their scrutiny is diverted away from those humans that devised and fielded the AI. The thing is, legally, in today’s world, there isn’t any legal recourse to somehow go after the AI for your efforts of righting an injustice when you’ve been wronged by an AI system. If you decide to legally pursue your just compensation for the injury by going after the AI, please realize that you aren’t going to be able to compel the AI to walk, crawl, or somehow enter into a courtroom to face the truth piercing scrutiny of justice.Instead, in today’s world, you could potentially pursue the hope for justice by going after the humans that underlie the AI system, including specific humans or possibly the company or companies that pertain to the AI system. They are traditionally within the sphere of being held legally responsible or legally accountable.That being said, remember that we are also now pretending that we are going to provide **legal personhood to AI or robots**, and if so, the scenario that I have sketched is going to radically change. The scenario won’t change, but the pointing of fingers is going to legally demonstratively change. Presumably, an AI that has **legal personhood** can be pursued for legal redress. You could try suing the AI directly for compensation regarding your **injuries.** Besides civil legal actions, you could possibly have the government seek criminal charges against the AI. For those of you with a curiosity about this sort of thing, you might take a moment to consider how the AI would be criminally imprisoned or otherwise judicially punished for its crime, assuming that the AI or robot was found guilty (and was not able to overturn the verdict on appeal). This determined **pursuit** of the AI or robot as to its legal responsibility due to having **legal personhood** seems perhaps judiciously satisfying. If it was the AI’s fault, and since it has the legal duty of care as presumably imposed via having legal personhood, the AI ought to pay the consequences for its wayward acts. No need to waste time going after any humans that were involved in the casting of the AI. Set your eyes on the AI and drag that AI into court. I assure you that coping with AI that has legal personhood and seeking redress from that AI is going to be a **rat’s nest** of legal challenges. I realize you might snicker at this next comment, but we would almost decidedly have **lawyers** that would be asked to represent the **AI or robot** during any justice-seeking endeavors. One would assume that legal personhood should also have adequate **legal representation**. Again, for a bit of an offshoot of mind-expanding considerations, if we can achieve AI-based legal reasoning sufficiently to perform the work of human lawyers, would we therefore possibly have AI that represents the AI that caused the injury? That’s a side tangent for you to ponder. Getting back to the matter of the **AI or robot** as having legal personhood and carrying the legal responsibility accordingly, suppose that the AI wasn’t truly at **fault**. Nonetheless, the humans that perhaps crafted faulty AI could hide behind the AI garnered legal personhood. Whereas today this kind of hiding is not going to get you very far, the gambit becomes much more viable if AI has legal personhood. At that juncture, the expectation is that the AI will be held legally accountable. Researchers at the University of Oxford stated the matter this way: “Attributing electronic personhood to robots risks misplacing moral responsibility, causal accountability, and legal liability regarding their mistakes and misuses. Robots could be blamed and punished instead of humans. And irresponsible people would dismiss the need for care in the engineering, marketing and use of robots” (by co-authors Luciano Floridi and Mariarosaria Taddeo, “Don’t Grant Robots Legal Personhood” in Nature, 2018). There are lots of **variations** concerning how this blame-shifting might take place. Someone might truly know that they were the bearers of the misconduct but opt to use the AI as a potential legal and ethical shield to protect their dour deeds. You can’t do so today because the AI of today does not have legal personhood. Once we allow **legal personhood** for AI or robots, the door is opened for this kind of shield-bearing shenanigans. Is it Pandora’s box that is being opened? Another possibility is that someone is unsure of whether they were the source of the **misconduct**, but they go along with the idea that it probably is the AI. This is an instance where the human is not trying to overtly pull a fast one. They are simply going to wait and see what happens when the pursuit of the AI or robot reveals whether the AI or robot is in fact the legally **accountable source** or not. Might true justice be delayed or obscured by this? As you can plainly see, the AI or robot will be a juicy scapegoat. A fiendish avenue also then becomes feasible. An **evildoer** embeds something untoward into an AI or robot system and then waits to see the desired adverse outcome. When the AI or robot causes injury or possibly fatalities, the AI or robot is the legal focus of the inquiries. It is conceivable that the wrongdoers might have time to escape or that they will never even be pursued due to an assumption or claim that it was only the AI or robot that was the perpetrator of the malicious act. Fiendish! Suppose evildoers merely tricked the AI or robot into performing a **monstrous act.** There might be little if any indication that the AI or robot was **swayed in that direction**. Once again, our attention anyway is first and foremost going to be on the AI since it has legal personhood as its banner. Without going wildly overboard, imagine that the fiends were so sneaky that they planted evidence into the AI or robot that would cause a computer-focused forensic analysis by human forensic experts to incorrectly surmise that the problem was within the AI or robot and not by some other external means. I realize this seems like a plot from the ever-classic Columbo series, but the insidious schemes of evildoers should never be underestimated (Columbo abundantly taught us that). Before we go into the stratosphere on this, there is the other side of the coin that you can equally consider. Suppose the **AI is so clever** that it can make things seem as though humans were legally responsible, trying to shift the **legal burden away** from itself. If that seems preposterous, remember that we are pretending that AI is good enough to garner legal personhood. We might reasonably expect that anything or anyone granted legal personhood is likely also clever enough to want to slide out of legal responsibility if feasible to do so, especially assuming that there are legal repercussions for getting caught. The usual retort is that we should program all AI to strictly speaking legally uphold its legal accountability and not weasel out of being legally ensnared. Well, all I can say is good luck with that. Round and round we go. Overall, we can recap that AI legal personhood might be: 1) No legal personhood for AI (today’s prevailing circumstance), or 2) Legal personhood for AI (with full legal accountability), or 3) A variant of legal personhood for AI such as limited or no legal responsibility Go ahead and vote for the choice of AI legal personhood that you think is best. Come back to this in five years, ten years, and fifty years, and see if your preferences have changed. I realize this has been a somewhat heady examination of the topic and you might be hankering for some day-to-day examples. There is a special and assuredly popular set of examples that are close to my heart. You see, in my capacity as an expert on AI including the ethical and legal ramifications, I am frequently asked to identify realistic examples that showcase AI Ethics dilemmas so that the somewhat theoretical nature of the topic can be more readily grasped. One of the most evocative areas that vividly presents this ethical AI quandary is the advent of AI-based true self-driving cars. This will serve as a handy use case or exemplar for ample discussion on the topic. Here’s then a noteworthy question that is worth contemplating: Does the advent of AI-based true self-driving cars illuminate anything about the AI legal personhood topic, and if so, what does this inform us about? Allow me a moment to unpack the question.First, note that there isn’t a human driver involved in a true self-driving car. Keep in mind that true self-driving cars are driven via an AI driving system. There isn’t a need for a human driver at the wheel, nor is there a provision for a human to drive the vehicle. For my extensive and ongoing coverage of Autonomous Vehicles (AVs) and especially self-driving cars, see the link here. I’d like to further clarify what is meant when I refer to true self-driving cars. Understanding The Levels Of Self-Driving Cars As a clarification, true self-driving cars are ones that the AI drives the car entirely on its own and there isn’t any human assistance during the driving task. These driverless vehicles are considered Level 4 and Level 5 (see my explanation at this link here), while a car that requires a human driver to co-share the driving effort is usually considered at Level 2 or Level 3. The cars that co-share the driving task are described as being semi-autonomous, and typically contain a variety of automated add-ons that are referred to as ADAS (Advanced Driver-Assistance Systems). There is not yet a true self-driving car at Level 5, which we don’t yet even know if this will be possible to achieve, and nor how long it will take to get there. Meanwhile, the Level 4 efforts are gradually trying to get some traction by undergoing very narrow and selective public roadway trials, though there is controversy over whether this testing should be allowed per se (we are all life-or-death guinea pigs in an experiment taking place on our highways and byways, some contend, see my coverage at this link here). Since semi-autonomous cars require a human driver, the adoption of those types of cars won’t be markedly different than driving conventional vehicles, so there’s not much new per se to cover about them on this topic (though, as you’ll see in a moment, the points next made are generally applicable). For semi-autonomous cars, it is important that the public needs to be forewarned about a disturbing aspect that’s been arising lately, namely that despite those human drivers that keep posting videos of themselves falling asleep at the wheel of a Level 2 or Level 3 car, we all need to avoid being misled into believing that the driver can take away their attention from the driving task while driving a semi-autonomous car. You are the responsible party for the driving actions of the vehicle, regardless of how much automation might be tossed into a Level 2 or Level 3. Self-Driving Cars And AI Legal Personhood For Level 4 and Level 5 true self-driving vehicles, there won’t be a human driver involved in the driving task. All occupants will be passengers. The AI is doing the driving. One aspect to immediately discuss entails the fact that the AI involved in today’s AI driving systems is not sentient. In other words, the AI is altogether a collective of computer-based programming and algorithms, and most assuredly not able to reason in the same manner that humans can. Why is this added emphasis about the AI not being sentient? Because I want to underscore that when discussing the role of the AI driving system, I am not ascribing human qualities to the AI. Please be aware that there is an ongoing and dangerous tendency these days to anthropomorphize AI. In essence, people are assigning human-like sentience to today’s AI, despite the undeniable and inarguable fact that no such AI exists as yet. With that clarification, you can envision that the AI driving system won’t natively somehow “know” about the facets of driving. Driving and all that it entails will need to be programmed as part of the hardware and software of the self-driving car. Let’s dive into the myriad of aspects that come to play on this topic. First, it is important to realize that not all AI self-driving cars are the same. Each automaker and self-driving tech firm is taking its approach to devising self-driving cars. As such, it is difficult to make sweeping statements about what AI driving systems will do or not do. Furthermore, whenever stating that an AI driving system doesn’t do some particular thing, this can, later on, be overtaken by developers that in fact program the computer to do that very thing. Step by step, AI driving systems are being gradually improved and extended. An existing limitation today might no longer exist in a future iteration or version of the system. I trust that provides a sufficient litany of caveats to underlie what I am about to relate. We are primed now to do a deep dive into self-driving cars and ethical AI questions entailing the AI legal personhood conundrum. Let’s use a readily straightforward example. An AI-based self-driving car is underway on your neighborhood streets and seems to be driving safely. At first, you had given special attention to each time that you managed to catch a glimpse of the self-driving car. The autonomous vehicle stood out with its rack of electronic sensors that included video cameras, radar units, LIDAR devices, and the like. After many weeks of the self-driving car cruising around your community, you now barely notice it. As far as you are concerned, it is merely another car on the already busy public roadways. Lest you think it is impossible or implausible to become familiar with seeing self-driving cars, I’ve written frequently about how the locales that are within the scope of self-driving car tryouts have gradually gotten used to seeing the spruced-up vehicles, see my analysis at this link here. Many of the locals eventually shifted from mouth-gaping rapt gawking to now emitting an expansive yawn of boredom to witness those meandering self-driving cars. Probably the main reason right now that they might notice the autonomous vehicles is because of the irritation and exasperation factor. The by-the-book AI driving systems make sure the cars are obeying all speed limits and rules of the road. For hectic human drivers in their traditional human-driven cars, you get irked at times when stuck behind the strictly law-abiding AI-based self-driving cars. That’s something we might all need to get accustomed to, rightfully or wrongly. Back to our tale. One day, suppose a self-driving car in your town or city suddenly rams into a pedestrian that was crossing the street. Sadly, the pedestrian was seriously injured (fortunately, they will recover to full health). Everyone goes up in arms over the self-driving car colliding with a pedestrian. Outrageous! A danger to society. A menace on our peaceful streets. Who is to blame for this reprehensible act? If this happened today (of which, similar cases have occurred), the knee-jerk reaction might be to blame the AI. The AI did it. No other explanation is possible. Unless a remote human operator was somehow intervening with the driving controls, the culprit has to be the AI driving system. Case closed. But, wait for a second, remember that the AI of today does not have legal personhood. You can certainly argue that AI driving likely is the root of what caused the collision. You can dig into the AI system and try to trace what happened as per the programming of the AI. All of that will aid in revealing what the AI did and did not do, which presumably led to the collision. Such a detailed technological inspection and review will be essential fodder for going after the AI developers, the automaker, the self-driving car systems builders, the fleet operators, and other human or human-based companies that had a hand in the self-driving car. We might also see a legal pursuit of the city or governmental body legally responsible for the stretch of road that the incident occurred at. You are not going to see anyone seriously try to legally pursue the AI per se. With that handy exemplar in hand, shift your mindset to the future. The future might consist of assigning legal personhood to various AI systems. Pretend that the AI driving system of this brand of self-driving cars had been legally declared as officially embodying legal personhood. The AI is now fair game for the legal accountability search mission. Let’s revisit the discussion earlier about ways in which AI might be utilized as a scapegoat, either by intention or by happenstance. We have these types of possibilities: Doer of misconduct hiding behind the AI as a shield for legal misdirection Potential misconduct by a human but waiting to see how the AI first fares Evildoer that implanted evil into the AI and hopes no one will be the wiser Wrongdoer that sneakily tricked the AI into performing an untoward act Bad doer that directly did the foul deed (not by the AI) while making AI the fall guy Etc. We can briefly apply each of those scenarios to the imagined incident of the self-driving car that has struck a pedestrian (we are assuming too that the pedestrian was not at fault or might only be partially at fault, thus the AI is still on the hook). Here’s how it might go: Doer of misconduct hiding behind the AI as a shield for legal misdirection An AI developer that crafted the AI driving system looks at their code and realizes that they mistakenly neglected to include a segment that would have prevented the collision from occurring. Yikes! The AI developer doesn’t want to lose their job. Their career is also on the line. When asked what happened, the AI developer explicitly states that it must be the fault of the AI and the AI ought to take the fall. Potential misconduct by a human but waiting to see how the AI first fares The company that is operating the fleet of self-driving cars is unsure whether they might have caused the collision or whether the AI did it. There is a chance that poor maintenance of the autonomous vehicle could account for why the collision occurred. On the other hand, it seems that the maintenance records showcase that the proper upkeep was taking place. The top executives of the firm agree to wait and see whether the AI gets legally pinned for the collision. Evildoer that implanted evil into the AI and hopes no one will be the wiser A hacker hitman was hired to injure a specifically targeted person that lived in the town and was known for frequently walking from their office to a nearby diner at lunchtime. The hacker used various cyber breaching tools to get a snippet of code into the AI driving system that would detect the specific person and attempt to ram into them. Sure enough, by routing the self-driving car to the known location and at the right time of day that the target routinely crossed the street, the evildoer code insertion directed the AI driving system to strike the person upon detecting their presence. It will be extremely hard for anyone to figure out that the AI evildoing took place since the snippet of code erased itself immediately after the mission was accomplished. Wrongdoer that sneakily tricked the AI into performing an untoward act A jilted lover of the pedestrian had decided that some revenge was warranted. They stood at the corner where the target routinely crossed the street just at sunset after work was over. Holding a well-polished mirror, the wrongdoer angled the mirror to relay the remaining sunlight into the cameras on the self-driving car. For a brief instant, the AI driving system was not getting needed sensory data and inadvertently plowed into the pedestrian. You could say that the wrongdoer tricked the AI by exploiting a known weakness of this particular AI driving system and its set of electronic sensors. Bad doer that directly did the foul deed (not by the AI) while making AI the fall guy A malcontent was standing next to the pedestrian. They did not know each other. Nonetheless, the troublemaker did not like the person. When a self-driving car came down the street, the bad doer pushed the person onto the street. The AI driving system could not stop the autonomous vehicle in time. Bam, the self-driving car struck the pedestrian and produced the corresponding injuries. The bad doer sprinted away to avoid being spotted and figured that the AI would be the fall guy for the incident.Conclusion For any of you that might quibble with the admittedly plotted examples, I believe they are within the bounds of reasonably being possible. I realize that many in the self-driving car industry are insisting that for example, no one will ever be able to cyber breach a self-driving car, but this is wishful thinking. You can bet your bottom dollar on cybercriminals and bad hat actors striking the myriad of AI driving systems and self-driving cars. It is only a matter of time until enough self-driving cars are on the roadways and whether there is any profitable or beneficial reason to hack them. See my further analysis on this at the link here. The primary overarching point is that we need to carefully consider whether we want to change our laws to include AI or robots as being able to possess legal personhood. This is a humongous question. Legal entanglements of immense proportions will ensue. Ethical entanglements of immense proportions will ensue. The acclaimed Roman statesman and lawyer Marcus Tullius Cicero had famously uttered this sage wisdom: “The safety of the people shall be the highest law.” We must ask ourselves, are we going to be safer or less safe if we opt to designate legal personhood to AI or robots? The good news is that you get to decide since the matter is still being debated and the die has not yet been cast. May the laws and our ethics do us well.

### AT: Aschenbrenner

#### Aschenbrenner is wrong

**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

Since then, the hype has continued to balloon. Leopold **Aschenbrenner**, who at the time was a researcher at OpenAI focusing on **superintelligence**, told me last year: “**AI progress** in the last few years has been just **extraordinarily rapid**. We’ve been crushing all the benchmarks, and that progress is continuing unabated. But it won’t stop there. We’re going to have superhuman models, models that are much smarter than us.” (He was fired from OpenAI in April because, he claims, he raised security concerns about the tech he was building and “ruffled some feathers.” He has since set up a Silicon Valley investment fund.)

In June, Aschenbrenner put out a 165-page manifesto arguing that AI will outpace college graduates by “2025/2026” and that “we will have superintelligence, in the true sense of the word” by the end of the decade. But others in the industry scoff at such claims. When Aschenbrenner tweeted a chart to show how fast he thought AI would continue to improve given how fast it had improved in last few years, the tech investor Christian Keil replied that by the same logic, his baby son, who had doubled in size since he was born, would weigh 7.5 trillion tons by the time he was 10.

It’s no surprise that “sparks of AGI” has also become a byword for over-the-top buzz. “I think they got carried away,” says Marcus, speaking about the Microsoft team. “They got excited, like ‘Hey, we found something! This is amazing!’ They didn’t vet it with the scientific community.” Bender refers to the Sparks paper as a “fan fiction novella.”

Not only was it provocative to claim that GPT-4’s behavior showed signs of AGI, but Microsoft, which uses GPT-4 in its own products, has a clear interest in promoting the capabilities of the technology. “This document is marketing fluff masquerading as research,” one tech COO posted on LinkedIn.

### AT: Healthcare AI

#### Pro-AGI healthcare research is vacuous rubbish and ignores risks of un-Hippocratic actions like eugenics from perverse AGI incentives

**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

Limitations of the **AGI literature** and future directions In the reviewed articles, there was a **limited number of studies** that focused on the risks to specific **doma**ins, as well as descriptions of specific functionality of the AGIs. The majority of studies reviewed were non-domain specific and focused on the general risks to humanity. However, specific domains identified in the eligible articles included autonomous vehicles, human professions, and smart homes (Chen & Lee, Citation2019), manufacturing, communication, and energy (Narain et al., Citation2019), law (Nindler, Citation2019), environment and social aspects (Pueyo, Citation2018). Two notable **domains** that did not feature in the review were research investigating the risks associated with defence and autonomous weapons systems, and **healthcare.** Given the obvious catastrophic risks associated with losing control of autonomous weapons systems to an AGI, an understanding of the potential risks is critical. AGI research with connections to the military is being conducted, in a 2017 survey of active AGI research and development projects, nine out of the total 45 active AGI projects had military connections (Baum, Citation2017). However, no military research was identified by the current review search strategy, which indicates two logical perspectives 1) the research may be confidential and not published, or 2) the research does not focus specifically on risk (Baum, Citation2017). In healthcare, we have witnessed how rapidly ANI has changed medical practice, for example, disease diagnosis, robotic surgery, and drug discovery, among others (Yu et al., Citation2018). Despite no inclusion of healthcare specific research in the current review, it is logical to assume that AGI systems in healthcare are being considered and may have enormous benefits. Baum (Citation2017) identified 20 active AGI projects with the stated goal being Humanitarian, which may include healthcare, yet was not explicitly stated. As with defence, there would be significant risks associated with AGI systems which have the capacity to make life or death decisions. Moreover, many philosophical discussions focus on healthcare AGI which may seek to optimise achievement of its stated goals via means which provide risk to human life. For example, **AGI system**s which are tasked with eradicating **diseases** such as cancer which establish that they can achieve their goal more efficiently by eradicating those in the population who have a genetic predisposition to cancer (Salmon, Hulme et al., Citation2021) or at the expense of other, longer-term **chronic conditions**. As such, research on the potential risks associated with the unsafe AGI, or the definition of ‘safety’ in the healthcare system is necessary future research direction. The review revealed that a majority of articlesdo not provide details of the **AGI system’s specifications** or functionality. Rather, most of the articles refer to generic AGI systems without describing what capabilities they possess, what goals they may have, and what tasks they will likely perform. It is our view that this is a significant limitation, as it impacts the quality of subsequent risk assessment efforts. Whilst many formal risk assessment methods exist (Dallat, Salmon & Goode Citation2019), most require at least some description of the tasks being performed and the goals being pursued, and state-of-the-art methods also require a description of the system in which the tasks are being performed (e.g., Dallat et al., Citation2018; Leveson, Citation2011; Stanton & Harvey, Citation2017). Without identifying the specifications of **different AGI** systems, it is not possible to accurately forecast the range of risks associated with them. Currently there are few modelling efforts investigating the diffusion of AGI into society (Narain et al., Citation2019), which was exemplified in the current review by the limited number of modelling approaches identified. Although our understanding of the risk associated with AGI is limited, they have similar characteristics to **other risks** that involve integration of humans and technology, and modelling techniques exist that may provide meaningful analyses in relation to AGI (Barrett & Baum, Citation2017). That said, a limitation of the current literature is that there are few studies employing formal scientific analysis methods to identify and assess the risks associated with AGI. Further research exploring the use of modelling approaches such as computational modelling (Salmon et al, Citation2020), systems analysis (Stanton et al., Citation2013), and risk assessment methods (Dallat et al., Citation2018) is recommended.

#### LLM in healthcare causes biases, misinformation and hallucinations

**Lalande 24** [Edgar Lalande is a journalist at Charlie Hebdo, 08-07-2024, "AI and medicine: the big scam", Charlie Hebdo,https://web.archive.org/web/20240324052423/https://charliehebdo.fr/2024/02/sciences/sante/ia-et-medecine-la-grande-arnaque/

\*note: article automatically translated from French

If you thought that **artificial intelligence** (AI) had the potential to improve the efficiency of **medicine** on a large scale , you were wrong. According to the latest report from the World Health Organization (WHO), the massive use of technologies based on **generative AI** in the **medical field** risks not only amplifying **social inequalities** between countries, but also leading to a “collapse of the health model” on a global scale. Unlike early AI algorithms, which simply stored databases to perform automated tasks, generative AI like that used by ChatGPT is capable of generating new content—text, images, videos—regardless of the formats of the input data. Adopted “faster than any other consumer application in history ,” generative AI is colonizing the healthcare field to produce clinical notes, help doctors diagnose and treat patients, and design new drugs. Biases without safeguards The problem, **WHO** points out, is that the **large companies** that develop these tools are mainly located in **rich countries**, and primarily target their populations. This **exacerbates** the **learning biases** of algorithms, which ignore a large part of the scientific data obtained from disadvantaged populations. However, many pathologies are multifactorial and depend on both genetic and environmental specificities. How can we hope to diagnose and treat melanomas if the algorithm has only trained on a skin type that evolves under certain latitudes? Even more seriously, the **WHO** fears a form of **race** to **mediocrity** on the part of **multinationals** engaged in an intractable economic competition. By wanting to put their products on the market as quickly as possible for essentially financial reasons, these companies will produce inaccurate, incomplete, and even dangerous results for **health.** For example, **algorithms** can "invent" false references leading to erroneous answers. Poetically described as "hallucinations" by researchers, these computer bugs are so well constructed that it is impossible to distinguish them from the **factually accurate answers** generated by the machine. Algorithms "are not trained to produce facts, but to produce information that looks like facts" , recalls the WHO. **Generative AI tools** "do not understand what they are "**saying**" and do not have the slightest moral or contextual reasoning to answer questions" . One study found that when given a simple set of facts to summarize, large language models "**hallucinated**" at least 3% of the time... sometimes as much as 27%! Without international safeguards, the situation could quickly get out of **control**. Because once released on the Web, this false medical information will circulate with the same credibility as that resulting from scientific consensus. A scenario of **disinformation** that will not be able to be stopped, which makes the **WHO** fear a "**collapse** of the model" of health in the world. Here again, it is the most disadvantaged populations who will suffer first. Often facing a shortage of doctors, they are the most likely to blindly follow medical information gleaned from the Internet. Information that is false or unsuitable for their specificities. Double jeopardy.

**AT: Securitization Kritik**

#### Err aff on AI threats - best studies, expert opinion, and magnitude of true existential risks

**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

For those who think worrying about AGI is falling for Pascal’s wager  Pascal’s wager is that given the infinite losses (hell vs paradise) incurred if we wrongly choose to not believe in God, we should act (wager) under the belief that God (the Christian god, by the way) exists. The argument against doing something about the catastrophic risks of AI draws the analogy to Pascal’s wager because of the huge risks, even potentially infinite if you consider the extinction of humanity that way. In the limit of infinite losses under extinction, we would have to act as if those risks are real with an amount of evidence or a **probability of extinction** that is allowed to go to zero (because the risk can be measured, in expectation, by the product of the probability of the event times the loss if it happens). Let us now see where that argument breaks down, mostly because we are not dealing with tiny probabilities. In a December 2023 survey, the median AI (not safety) researcher put 5% on AI causing extinction-level harm. Probabilities of 5% are not Pascal’s wager territory. There are serious arguments supported in the scientific literature (see “the report” and a lot of the discussion above) for the various kinds of catastrophic risks associated with very advanced AI, especially as we approach or surpass human level in some domains. Also, we do not need to take the losses to infinity: There are many potentially very harmful possibilities along the path to AGI and beyond (again, see “the report”). So we end up with non-zero evidence for AI catastrophes and the possibility of non-infinite but unacceptable losses, the usual setting for decision theory, and rationality thus demands that we pay attention to these risks and try to understand and mitigate them.  For those who discard x-risk for lack of reliable quantifiable predictions   Of course no one has quantitative models of future scientific advances along with social and political change regarding AI. Hence we cannot run quantitative models such as those applied to sample future climates. The only quantitative options are individual and aggregated subjective probabilities, e.g., from polling experts. Can we trust the 5% median x-risk in this recent study? I would say to some extent, in ways similar to how we can trust the aggregate long-term predictions of economists. Are they sufficient to drive policy? No, not alone, but they surely send an important signal because experts internalize their understanding of the world and apply their system 1 computation to it, i.e., intuition, in ways that can be very valuable. But of course, and very importantly, we also need to consider rational but not fully quantitative arguments such as those I outlined above. For example, we can ask questions like, “what if we build a superintelligent AI, and what if it has goals that are dangerous to humanity?”. There are many ways in which one can argue that superintelligence is plausible (with lots of uncertainty about the timeline) and many ways that have been discussed in which an AI acquires dangerous goals, the simplest being that a human provides them. Should this uncertainty make one conclude that public policy should not consider AI x-risk? Of course not; given the magnitude of the potentially negative impact (up to human extinction), it is imperative to invest more in both understanding and quantifying the risks and developing mitigating solutions. And the uncertainty in timeline means that, yes, there is urgency in doing these things, in case AGI happens faster than expected.

#### Best insider and expert assessments verify the realism of AI threats

**Perrigo 24** [Billy Perrigo, correspondent for TIME, shortlisted for the 2022 Orwell Prize, 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.

**AT: Fix Alignment CP**

#### Alignment fails – tech, speed, novelty, and geopolitics

**Leahy et al. 24** [Connor Leahy, artificial intelligence researcher with a degree in computer science from the Technical University of Munich, Gabriel Alfour, artificial intelligence researcher and CTO at Conjecture, an AI safety research company, Chris Scammell, artificial intelligence researcher with a BA in computer science from Colby College, Andrea Miotti, tech consultant with Double Master's Degree in Public Policy from the The London School of Economics and Political Science (LSE) , and Adam Shimi, PhD student in theoretical computer science at the L'Institut de recherche en informatique de Toulouse, 12-09-2024, "AI Safety", The Compendium, https://www.thecompendium.ai/ai-safety]/Kankee

Today’s AI safety research is vastly underfunded compared to investments in capabilities work, and the majority of technical approaches intentionally do not address the conceptual complexity of alignment, instead operating in a reactive  empiricist framework that simply identifies misbehavior once it already exists. Humanity’s current AI safety plan is to race toward building superintelligent AI, and delegate the most difficult questions of alignment to AI itself. This is a naive and dangerous approach. For a safe future, we must solve the hard problems of alignment, allocating adequate research hours, investment, and coordination effort. OpenAI, Deepmind, Anthropic, X.AI (“accelerating human scientific discovery”), and others have all proposed deferring and outsourcing these questions to more advanced future AI systems. But on reflection, this is an **incredibly risky** approach. Situational Awareness, a document written by ex-OpenAI superalignment researcher Leopold Aschenbrenner which has gotten significant traction even from popular news outlets, puts the argument bluntly. Aschenbrenner argues for a vision of the future in which AI becomes powerful extremely quickly due to scaling up the orders of magnitude (“OOMs”) of AI models. When discussing future safety approaches, he makes a vivid argument for iterative alignment: "Ultimately, we’re going to need to automate alignment research. There’s no way we’ll manage to solve alignment for true superintelligence directly; covering that vast of an intelligence gap seems extremely challenging. Moreover, by the end of the intelligence explosion—after 100 million automated AI researchers have furiously powered through a decade of ML progress—I expect much more alien systems in terms of architecture and algorithms compared to current system (with potentially less benign properties, e.g. on legibility of CoT, generalization properties, or the severity of misalignment induced by training). But we also don’t have to solve this problem just on our own. If we manage to align somewhat-superhuman systems enough to trust them, we’ll be in an incredible position: we’ll have millions of automated AI researchers, smarter than the best AI researchers, at our disposal. Leveraging these army of automated researchers properly to solve alignment for even-more superhuman systems will be decisive. Getting automated alignment right during the intelligence explosion will be extraordinarily high-stakes: we’ll be going through many years of AI advances in mere months, with little human-time to make the right decisions, and we’ll start entering territory where alignment failures could be catastrophic." The dangers here are explicit: alien systems, huge advances in mere months, and a tightrope walk through an “intelligence explosion” in which wrong choices could lead to catastrophe.  But even before we get to a dramatic vision of the AI future, the iterative alignment strategy has an ordering error – we first need to achieve alignment to safely and effectively leverage AIs.  Consider a situation where AI systems go off and “do research on alignment” for a while, simulating tens of years of human research work. The problem then becomes: how do we check that the research is indeed correct, and not wrong, misguided, or even deceptive? We can’t just assume this is the case, because the only way to fully trust an AI system is if we’d already solved alignment, and knew that it was acting in our best interest at the deepest level. Thus we need to have humans validate the research. That is, even automated research runs into a bottleneck of human comprehension and supervision. Proponents of iterated alignment argue that this is not a real issue, because “evaluation is easier than generation.” For example, Aschenbrenner further argues in Situational Awareness that: "We get some of the way [to superalignment] “for free,” because it’s easier for us to evaluate outputs (especially for egregious misbehaviors) than it is to generate them ourselves. For example, it takes me months or years of hard work to write a paper, but only a couple hours to tell if a paper someone has written is any good (though perhaps longer to catch fraud). We’ll have teams of expert humans spend a lot of time evaluating every RLHF example, and they’ll be able to “thumbs down” a lot of misbehavior even if the AI system is somewhat smarter than them. That said, this will only take us so far (GPT-2 or even GPT-3 couldn’t detect nefarious GPT-4 reliably, even though evaluation is easier than generation!)"The argument holds for standard peer-review, where the authors and reviewers are generally on the same intellectual level, with sensibly similar cognitive architecture, education, and knowledge. But this does not apply to automated alignment research, where to be useful the research needs to be done by AIs that are both smarter and faster than humans. The appropriate analogy is not one researcher reviewing another, but rather a group of preschoolers reviewing the work of a million Einsteins. It might be easier and faster than doing the research itself, but it will still take years and years of effort and verification to check any single breakthrough. Fundamentally, the problem with iterative alignment is that it never pays the cost of alignment. Somewhere along the story, alignment gets implicitly solved – yet no one ever proposes an actual plan for doing so beyond “the (unaligned) AIs will help us”. There are other risks with this approach as well.  The more powerful AI we have, the faster things will go. As AI systems improve and automate their own learning, AGI will be able to improve faster than our current research, and ASI will be able to improve faster than humanity can do science. The dynamics of intelligence growth means that it is possible for an ASI “about as smart as humanity” to move to “beyond all human scientific frontiers” on the order of weeks or months. While the change is most dramatic with more advanced systems, as soon as we have AGI we enter a world where things begin to move much quicker, forcing us to solve alignment much faster than in a pre-AGI world. **'Tensions** between world powers will also heat up as AI becomes more powerful, something we are already witnessing in AI weapons used in warfare, global disinformation campaigns, the US-China chip war, and how Europe is struggling with regulation around Big Tech. As we move towards AGI, ASI, and eventually godlike AI, pressure on existing international treaties and diplomacy methods will be pushed beyond their limits. Unlike with nuclear war, there is not necessarily the same promise of mutually assured destruction with AI that could create a (semi)stable equilibrium. Ensuring geopolitical stability is necessary to create supportive conditions to solve the hard problems of alignment, something that gets more challenging if AI is becoming rapidly more powerful.  AGI and its successor AIs will also cause massive political, economical, and societal destabilization through automating disinformation and online manipulation, job automation, and other shifts that look like “issues seen today but magnified as systems grow stronger”. This in turn makes coordination around massive research projects like the ones necessary to solve alignment extremely difficult.  Thus, iterative alignment fails on multiple accounts. In addition to not addressing the hard parts of alignment, it also encourages entering a time-pressured and precarious world. We have seen that alignment is an incredibly complex technical and social problem, one of the most complex any civilization needs to handle. And while the costs are enormous, no one is even starting to pay them, instead hoping that they will disappear by themselves as AIs become more powerful. In light of this failure to address the risks of godlike-AI from a research angle, it’s necessary to aggressively slow down and regulate AI progress, in order to avoid the catastrophe ahead. This comes from strong AI regulations, policies, and institutions. Unfortunately, as we explore next, the landscape is as barren here as it is on the research side.

#### Confinement fails - oracle AI

**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

4.2 Risks of AI from treacherous turn and before it reaches the “wild” It appears that there is not much risk from AI before it leaves its initial confinement (goes into the “wild”).  However, it still can give bad advice or use other thin information channels (e.g., text interfaces) to create damage outside and increase its own chances of freedom. For example, an oracle AI may be limited to giving short text advice via a very simple interface. But such advice, while seemingly benefcial to humans, may have subtle remote consequences, resulting in the liberation of, and an increase in, the power of the oracle AI (Bostrom 2014). Stanislav Lem wrote about the risks of oracle AI in his book “Summa Technologia” (Lem 1963). Such AI may give advice that appears to be good in the short term, but its longterm consequences could be **catastrophic**. In Lem’s example, the oracle AI advises humans to use a specifc type of toothpaste and, separately, a specifc type of anti-baldness treatment. These activate two genes, which are dangerous only in combination. Moreover, the AI did not do it because it had malevolent intent to exterminate humanity, but because it just searched for the best solution for a given goal among many options. However, the goal that humans gave to the AI in Lem’s example is dangerous: stop population increase. AI could stage a global catastrophe of any scale to facilitate its initial breakout from its creators. For example, it could stage a nuclear war, so that its operators release it into the wild, hoping that it will help them in the war. The AI could then create a global risk and demand full power, rightfully claiming that only it could prevent the risk. An AI may also falsely predict an **impeding risk** and demand to be released from confnement to prevent the risk. 4.3 AI risks after it leaves initial confinement but before it takes over the world

**AT: UBI Capitalism CP/DA**

#### UBI fails to solve capitalism – it doesn’t alter class dynamics

**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 COMMUNIST ORIENTATION TO AI Left accelerationism’s blind spot is reflected in a series of proposals that would, we think, do more to accelerate AI-capital than outdistance it with AI-socialism. These include support for a universal or guaranteed income as an answer to technological job loss; neglect of the ecological problems of intensive AI use; and a confidence in a pacific transition to socialism that overlooks the military and repressive aspects of AI. What is critical to a communist orientation to AI is the issue of the ownership and control of the means of production, a point that postcapitalist and left-accelerationist thinkers partially take up, but also obscure by their insistence on the possibilities of passage to high-technology socialism by parliamentary reforms undertaken within the framework of actually-existing AI-capitalism. We address these issues in turn. The ‘AI Plus UBI’ Formula Proposals for a universal basic income (UBI) as a transitionary component of an anti-capitalist programme have been discussed for many years (and supported by one of the authors of this book). One conclusion of these debates is that UBI’s political valence, whether as a permanent strike fund for labour or a streamlining of neoliberal welfare austerity, depends on the terms under which it might be instituted, e.g., generous or stingy, with or without a dismantling of other social benefits. The more basic issue, however, is that, introduced within capital, UBI does not disturb the ownership of the means of production, and in some ways endorses it, as a manifestation of the munificent largesse of the ruling class, offered within the context of an otherwise fully commodified economy (Clarke 2017). These issues have been recently highlighted by a sudden wave of enthusiasm for UBI among Silicon Valley capitalists, who are advancing the idea specifically as an antidote to the unemployment and precarity likely to be caused by AI. Facebook founder Chris Hughes (2018),  venture capitalist Marc Andreessen, and web guru Tim O’Reilly, among other Silicon Valley luminaries, support UBI as the ‘social vaccine of the 21st century’; the tech incubator Y Combinator is running a basic income research programme in Oakland; and tech entrepreneur Dan Yang has announced an independent US presidential candidacy on a platform that includes a form of basic income (Ghaffrey 2018; Ito 2018). These proposals tend to envisage UBI, often at a fairly low subsistence level, as an addition to an otherwise nakedly laissez-faire market order. As several observers (Filoux 2018; Sadowski 2018; Rushkoff 2018) have remarked, these proposals do not challenge the right of capital to direct AI development, **reaping billions**, and **do not disrupt** the vast **income inequalities**, either between capitalists and workers (or non-workers), or between elite professional high-tech employees and menial workers. And while UBI is promoted as an aid to entrepreneurialism, in reality it would probably require supplementation by forms of precarious work, in that respect actually subsidizing AI-driven gig-economy ventures such as Uber, Mechanical Turk or Figure Eight. Further, UBI makes no provision for other public supports, which might make the free time supported by UBI something other than a miserable penury. In this form, UBI figures as a holding pen for what Harari, whose gloomy futurism is favoured by many Silicon Valley capitalists, ruthlessly characterizes as a ‘useless class’ (2016: 379). This is certainly not the UBI that left accelerationists and postcapitalists want, but it is likely the type of UBI they would get under the auspices of AI-capital. In this regard we agree with the point made by  Alex Gourevitch and Lucas Stanczyk (2018): a basic income of some sort might be an important part of a mode of production beyond capital, but  it is not a prelude to it – rather, a trap waylaying its emergence. AI’s Dirty Secret

#### AI based UBI leads to symbolic violence

**Bélisle-Pipon 25** [Jean-Christophe Bélisle-Pipon, Assistant Professor at Simon Fraser University, Visiting Researcher at the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at the Harvard Law School, recieved a Bachelors of Science,Masters, and PhD from the University of Montreal, 02-24-2025, “AI, universal basic income, and power: symbolic violence in the tech elite's narrative”, Frontiers, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1488457/full]/Kankee

While economic, social, and normative analyses have been put forward in articles in Frontiers in Artificial Intelligence (Ernst, 2022; Huo et al., 2024; Merola, 2022), two key dimensions that remain underexplored in the UBI discussion are 1) the utilitarian calculation behind the AI-justified UBI narrative; and 2) the associated **symbolic violence**, as articulated by sociologist Pierre Bourdieu. I argue that UBI, while ostensibly a tool for social good, may end up justifying even greater disparities in wealth and entrench symbolic violence by **reinforcing divisions** between AI owners, those skilled or capacitated in using AI, and those who are merely recipients of its “benefits.” This symbolic violence is particularly perverse as it perpetuates a narrative of AI as universally beneficial, when in reality, it risks exacerbating socio-economic inequalities and creating profound epistemic and symbolic **injustices**. The AI elites' advocacy for UBI The advocacy for UBI by AI elites is a relatively new phenomenon. Figures like Elon Musk, the CEO of Tesla, SpaceX, and X, and Sam Altman, the CEO of OpenAI, have positioned themselves as champions of UBI. Musk (2024) recently indicated about the rise of AI that “In a benign scenario, probably none of us will have a job. There would be universal high income. There would be no shortage of goods and services. The question will really be one of meaning: if a computer can do, and the robots can do, everything better than you, does your life have meaning? I do think there's perhaps still a role for humans in that we may give AI meaning.” For his part, Altman (2016) indicated that “[he's] fairly confident that at some point in the future, as technology continues to eliminate traditional jobs and massive new wealth gets created, we're going to see some version of this at a national scale.” AI elites argue that as AI and automation increasingly replace human labor, UBI will be essential to prevent widespread economic dislocation and social unrest. This argument may be compelling, especially in a world where technological advancements threaten to render large segments of the workforce obsolete (Islam, 2024). However, the promotion of UBI by these tech magnates is not simply a philanthropic gesture; it is deeply intertwined with their interests in the expansion and dominance of AI technologies. Crane et al. (2019) argue that corporate strategies often align with maintaining and enhancing power structures that benefit corporate elites. The advocacy for UBI by AI leaders can be seen as a strategic move to pre-emptively address potential backlash against AI-induced risk and negative externalities, such as job losses or job polarization [i.e., reducing middle wages, shifting demand toward low and high wages (see Goos and Savona, 2024)], thereby securing a favorable business environment for continued AI development and deployment. Without going so far as to say that AI may be an existential risk (or X-Risk, a risk to the very viability of humanity)—as other movements such as the members of the effective altruism movement and the associated cause of longtermism may do (Jecker et al., 2024)—AI may pose significant economic and social risks if job losses are not offset. The narrative presented by these AI leaders suggests that UBI is a necessary adaptation to the inevitable rise of AI—a tool to ensure that everyone benefits from technological progress. Yet, this narrative serves to legitimize and reinforce the power dynamics that already exist in the AI industry. By advocating for UBI, these AI elites position themselves as benevolent visionaries who are concerned about the wellbeing of humanity. However, as Sadowski (2016) argues, promoting UBI can be a strategic way for AI elites to deflect criticism, maintaining control over narratives about AI's future while avoiding challenges to their profit motives. This framing distracts from the fact that the same individuals who are pushing for UBI are also those who stand to gain the most from the proliferation of AI technologies (Spencer, 2024). Bourdieu's concept of symbolic violence offers valuable insight into the deeper implications of UBI in the context of AI; however, it is important first to examine UBI from its utilitarian foundation. A utilitarian justification for UBI This narrative aligns with a utilitarian view for assessing the benefits and risks of AI in society. AI elites apply a utilitarian calculation, evaluating the moral justification of replacing humans with AIs by weighing the potential to maximize societal wellbeing against the associated harms. From this eudemonic standpoint—focused on balancing wellbeing and harm, or even more simply pleasures and pains—they envision a future where AI's dominance across human-dominated fields leads to a society characterized by widespread leisure and, for some, heightened performance. In their view, this transformation is morally defensible if measures (such as UBI) are implemented to mitigate the negative effects and ensure distribution of certain benefits for all. UBI is thus used to justify the possibility, and to demonstrate, that AI can provide for humanity's basic needs, while at the same time justifying that some can be ultra-wealthy and possess these technological tools of humanity's (apparent) sustenance (Islam, 2024). While it has not been directly invoked up to now, this is a curious application of John Rawls' principle of difference, which in his “Theory of Justice” states that social and economic inequalities are to be arranged so that they are to the greatest benefit of the least advantaged members of society, consistent with the just savings principle and the principle of fair equality of opportunity (Rawls, 1971). Arguably, providing UBI to all does not solve everything; it creates more equality amongst the less well-off, without acting to address inequalities and wealth gaps. Yes, this would be a first for humanity—an economic safe net from which all could benefit (which appears to be of a fixed-benefit nature, with no indication of adjustment to economic trends)—but this cannot justify the kind of leanness in which it seems to place the non-owners of AI (i.e., virtually the entire world population) compared to the AI elites. It is hard to make a convincing utilitarian claim that this is for the benefit of the less well-off. Furthermore, as Sen (2009) argues, a focus on utility maximization may neglect the distribution of capabilities and freedoms, which are essential for genuine social justice; something very plausible if like Musk (2024) is thinking that if “[computers and robots can be doing] everything better than [humans], does [human] life have meaning?” With AI potentially representing a X-risk (Jecker et al., 2024) or at the very least risking to lead to a “a shift in power toward actors with the capital and authority to deploy powerful AI systems, such as elites, corporations, and governments” (Dafoe, 2018), it is very unclear that AI will actually maximize utility and be for everyone's benefits. Unfortunately for AI-justified UBI proponents, a study funded by Altman has found that UBI is not a comprehensive solution to the economic challenges posed by AI-driven job loss (Ropek, 2024). The research, conducted by OpenResearch (2024) between 2020 and 2023, provided $1,000 a month to 1,000 low-income individuals, with a control group receiving $50 monthly. While UBI helped participants cover essential expenses like housing and groceries, it did not lead to significant improvements in employment quality, education, or overall health. The study concluded that while UBI can alleviate some immediate financial stress, it **falls short of addressing** deeper systemic **issues** such as healthcare access, job stability, and upward mobility. Thus, UBI alone is unlikely to mitigate the broader economic impacts of AI on the workforce. So, unfortunately, proponents' utilitarian calculation does not seem to be working as well as they would like. As a result, the impetus for supporting UBI seems more ideological and self-serving than beneficial. Symbolic violence in the AI-justified UBI narrative Here, Pierre Bourdieu's concept of “violence symbolique” can help deceive AI elites' benevolent narrative. Symbolic violence refers to a form of domination that is subtle and often imperceptible, yet profoundly effective in maintaining social hierarchies (Bourdieu and Wacquant, 1992). Symbolic violence operates through the imposition of meanings that are accepted as legitimate, even by those who are subordinated by them (Bourdieu, 1993). This form of violence is not physical, but it is deeply embedded in the social structures and cultural norms that shape our understanding of the world. With society's increased digitalization, Couldry and Mejias (2020) denote how data practices can constitute a new form of colonialism, reinforcing existing power structures through symbolic means. Such symbolic violence allows dominant groups perpetuate their power without overt coercion, by making their worldview appear natural and inevitable. In the context of UBI and AI, symbolic violence manifests in the way that the narrative of AI, as a universal good, is constructed and disseminated. The AI elites' promotion of UBI suggests that the best way to address the disruptions caused by AI is to provide people with a guaranteed basic income, thereby ensuring that everyone benefits from technological progress. However, this narrative obscures the deeper structural inequalities that are being reinforced by the same technologies that UBI is supposed to mitigate. Symbolic violence, Bourdieu and Wacquant (1992, p. 172) notes, “accomplishes itself through an act of cognition and of misrecognition that lies beyond—or beneath—the controls of consciousness and will.” In the case of AI-justified UBI, public's acceptance of this proposal as a universal good would be a form of misrecognition that a symbolic violence is being perpetrated and instead considering that AI and UBI are normal and rather logical within the existing social order. Such acceptance would be a legitimization of the power of the AI elite by presenting UBI as the solution to the very problems their technologies create, thus reinforcing the existing social order. AI-justified UBI as a tool for maintaining social hierarchies UBI, as promoted by AI elites, can be seen as a tool of symbolic violence in several ways. First, it reinforces the division between those who own and control AI technologies and those who are merely consumers of its benefits. The owners of AI—who are also the primary advocates of UBI—are positioned as the benevolent providers of a safety net for the masses. Meanwhile, the recipients of UBI are cast as passive beneficiaries of a system that they have little control over. This dynamic perpetuates the power of the AI elite, while simultaneously legitimizing their dominance by presenting them as the solution to the very problems that their technologies create. Moreover, UBI as a form of symbolic violence operates by masking the true nature of the inequalities that it purportedly seeks to address. By providing a basic income, the narrative suggests that the economic and social disruptions caused by AI can be managed and mitigated. However, this narrative ignores the fact that UBI does nothing to address the underlying power imbalances that give rise to these disruptions in the first place. Critics argue that UBI, without accompanying structural reforms, may fail to address underlying inequalities (Parijs and Vanderborght, 2017), just like the OpenResearch study hinted. As Jarow (2024) puts it, “hitching the case for basic income to fears of rapid AI progress makes it far more vulnerable than it needs to be.” By linking UBI to AI, its advocates risk creating a policy that merely manages the symptoms of economic inequality without addressing the root causes. This approach perpetuates a superficial solution that maintains the status quo, allowing the AI elite to continue accumulating wealth and power while the majority remains dependent on the systems that marginalize them. The perverse nature of symbolic violence in UBI The symbolic violence inherent in the promotion of UBI by AI elites is particularly perverse because it creates the **illusion** of inclusivity and fairness; values central to AI ethics (Victor et al., 2024). The narrative of UBI as a universal good suggests that everyone stands to gain from the increased presence of AI in our societies. However, this narrative obscures the fact that the benefits of AI are not distributed equally, and that UBI, as currently envisioned, may actually entrench existing inequalities rather than alleviate them. By framing UBI as a necessary response to AI-induced unemployment, the AI elites are effectively shifting the focus away from the need for more equitable distribution of power and resources. Those who control AI technologies continue to benefit disproportionately, while those who are dispossessed by these technologies are offered only a minimal safety net in return.

#### Techno-communism fails

**Butollo and Nuss 22** [Florian Butollo, Research Fellow for Globalisation, Work, and Production at the Weizenbaum Institute for the Networked Society, Berlin, Sabine Nuss, managing director of the Karl Dietz Verlag, political scientist, published books on Karl Marx's analysis of capitalism, economic crises in capitalism, digitalisation and automation, 02-20-2022, “Marx and the Robots Networked Production, AI and Human Labour”, Pluto, https://www.plutobooks.com/9780745344379/marx-and-the-robots/]/Kankee

Accumulation by innovation Rationalisation in production has always been the method of choice used by capital to edge out competitors. This encompasses both producing more cheaply and capturing market share through product innovation. Hence, competition is the driving force behind capitalism’s historically unprecedented power to innovate, to creatively destroy and to disrupt as desired. Mark Zuckerberg’s widely known motto, ‘move fast and break things’, could well have come from Schumpeter (‘creative destruction’) or Clayton Christensen (‘disruptive innovation’).25 Capitalism is the **most innovative** social order which humanity has brought forth thus far, and at the same time the **most destructive**: when people speak of the Anthropocene, the era of human mastery over the planet, they should actually speak of the Capitalocene.26 After all, only after industrialisation did humanity manage to permanently alter the earth’s surface. In the Marxian critique of political economy, innovation exhausts itself because it is external to the ‘normal’ production process; it does not initially play a role in the capitalist production process. Innovation by itself does not produce value and does not concern the labour theory of value. Here, innovation remains a special case, an exception, a coup scored by a single capital against the competition. Can innovation go from being the exception to the norm? Morris Suzuki puts forth the thesis that this can indeed happen given two conditions: the marginalisation of the significance of direct human labour input in the production process, simultaneously accompanied by the creation of information products becoming the central object of profit generation, in which case ‘innovation becomes the core  of the company’s profit-making activity’. Under these conditions, companies would be forced to shift ‘the center of gravity of surplus value creation […] away from the production of goods and towards the production of innovation – that is, of new knowledge for the making of goods’.27 The unchanging production of the same physical object is the norm in the production of material goods. The xth copy of a specific model of chair may well find a buyer, as concrete labour has been expended to produce said copy. With digital information, this is no longer the case. Once information has been created, it need not be produced from scratch again; as far as business is concerned, it has been burned. The consequence of this situation is the generation of an ever-changing product, a product which constantly and continuously modifies itself and thus always stays a step ahead of the competition. In Morris-Suzuki’s scenario, the exception becomes the rule as occasional leaps in innovation are replaced by a continuous ramp. Innovation takes on a new meaning: no longer just a weapon against the competition, it becomes the core of a company’s profitable activity. Morris-Suzuki refers to this principle as the ‘perpetual innovation economy’.28 Here, ‘surplus knowledge’ is central to the ‘incessant generation of new products and new methods of production’. 29 Perpetual innovation abolishes a linear accumulation model which continuously aggregates individual elements of labour as a labour product,  replacing it with a non-linear model. Increasingly often, repetitive labour processes are automated, almost without resource expenditure (such as with the streaming of digital content), and at zero marginal cost. AI-based products in particular are characterised by constant self-improvement. The South Korean electronics corporation LG has launched an entire product line around this characteristic: marketed under the heading ‘Evolve’, products are advertised which supposedly not only do not age, but remain forever young, changing with  and being educated by the user almost as if they were children being raised by parents. As LG’s Technical Director I. P. Park explains: ‘The age of artificial intelligence allows us to go even further. Information will now get better, richer, and deeper with usage and time.’30 the latest machine of capital, for now Ernst Bloch once remarked that it was not disputed in Marxist circles ‘that each latest machine which late-bourgeois technology produces is the best’.31 With good reason, artificial intelligence can be described as the latest machine of digital capitalism, at least for the time being. With its help, digital capitalism has succeeded in consolidating a new social operating mode in which the extraction, evaluation and valorisation of data has become the focus of economic activity. Given the current wave of new applications that are being labelled ‘Artificial Intelligence’, we are essentially dealing with data-driven software that can be successfully used in certain closely limited areas of application. We are witnessing a phase of AI development in which the technologies owned by the tech corporations have become mass marketable and thus turned into everyday products. Their function is limited to specific tasks such as the categorisation of images, the detection of patterns in large amounts of data, the representation of human–machine interfaces and the like. The construction of thinking machines is (still) not on the agenda – the current task at hand for the relevant actors is rather to consolidate the data-extracting business models. In many areas, user reactions are fed back into the system in realtime, i.e. without any noticeable delay between input and output, between action and reaction, which in turn contributes to optimising the system’s runtime. User signals, i.e. user behaviour patterns, are used for personalised responses, error recovery and the incremental product improvement in a cybernetic control loop. AI-driven services are optimised through the consumers themselves while being fed by their user data. All this occurs within a circuit of perpetual innovation with the simultaneous goals of optimisation and innovation as well as the monetisation of the consumers’ activities – with a constantly improving user experience. The user becomes a three-headed being: he or she is at once the customer, the supplier and the product. AI represents a technology that serves the aims of analysing Big Data, efficiently valorising the free labour of users and monitoring the digital workforce, just like the assembly line was (and continues to be) both a means of rationalisation and an instrument of power in  the factory. Artificial Intelligence is, at least for now, the latest innovation of digital capitalism, one through which it **seeks to expand** its globally dominant role in economy and society, consolidate its business models and secure its valorisation model. All three mechanisms of capital valorisation – the valorisation of general knowledge as a service, the exploitation of the gratuitous labour of users, and perpetual innovation as a source of profit – are realised in contemporary digital capitalism. Multiple forms of ‘a productive expenditure of human brains, muscles, nerves, hands [and hearts, T.D.]’32 beyond the wage labour form create labour products that are useful for capital. As a result, while actions and reactions are fed back into the system, all areas of life come to serve as a resource for generating profit.

### AT: AGI Impossible

#### AGI is inevitable from exponential computation and innovation growth. Err aff per psychological bias against understanding the pace of fast tech growth

**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, and Naz Shuva, lecturer at Columbia University with a bachelors from the New York Institute of Technology, 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

Introduction Humanity is heading towards unprecedented circumstances in a destined rendezvous of two distinct forces. The continuation of artificial intelligence (AI) development toward artificial general intelligence (AGI) at the current rate and direction without an increase in health measures informed by neuroscience has serious mental health implications for human beings. The meeting of the exponential nature of AGI and the ancient nature of human psychology and biology is likely to result in unrecoverable psychological and social damages, specifically because of A(G)I’s (AI and AGI) consequences on the human mind and our relational experiences. Considering the current trends of social media recommendation algorithms’(SM-RA) effects on the human brain, early observations from large language models (LLMs), and insights from neuroscience, we see a side of this trending set of reinforcing feedback loops amplifying internal human suffering. This distress could span as far as atrophy of higher functions of the brain and mind. Envision the timeline of humanity’s existence as a 1,000 page book. Based on what we know of human history, all of civilization would be contained in the final few pages. Much of the technological progress happened in the last few paragraphs[1]. Given that technology’s rate of change is exponential, in the next 25 years, we will see something like 100 years worth of progress. And in 100 years from now, we will see something like 20,000 years worth of change. The ancient nature of our minds and bodies, which still reflect basic survival instincts that were required for the earliest homo sapiens, cannot comprehend nor keep up with this speed of growth. While evolution can be hastened by external conditions, our attention, memory, and emotional networks cannot keep up with the steep curve of current technological development shy of human-machine integration.[2] If we disregard the impacts of this discrepancy in the creation phase of AGI, a dystopian reality may be bad enough to outweigh the utopia. A mismatch to our new psychological ecology (the digital world) would result in a chaotic subjective experience. People who create AI are generally not experts in mental health, and experts in mental health tend not to build tech. We cannot afford this division as this technology will affect all of civilization at a foundational level. A secure bridge between these disciplines in the planning and development of A(G)I is required for a humane future. Tools intended to serve humanity should be designed by humans for humans with the purpose of promoting well-being. Builders of technology prioritizing their enamor for inventiveness and efficiency over human happiness invites disastrous results for what we consider a fulfilling life. We need to alter the current course of development and consider how to best ensure that we remain in control of this technology with care for the human condition. An approach that favors the utopian fantasy of playing God and seeing what happens is reckless and potentially self-destructive. Instead, our innovation must be channeled through the collective wisdom accumulated through the ages about what it means to live a good life. Forces at Play Exponential Tech To draw a picture of the likely future, it is vital to appreciate the counterintuitive nature of exponential technological growth, as well as AGI as its driver. Considering Moore’s Law, we know that computational power doubles every two years[3]. That means by 2070, computational power will have doubled 23.5 times what it is right now. We’ve done the math. Computational power will be over 8 million times that which it is today. It is nearly impossible to imagine what this might mean for humanity. What will the world look like when there is a technology with that much power and probable decision-making capabilities? And what does the internal human experience feel like at this point - after 47 years of atrophy caused by machines solving our problems for us? The default, unworked mind is one that leads to suffering, as will be explained later. Kurzweil’s Law of Accelerating Returns expands Moore’s Law to other evolutionary systems and explains that the “rate of exponential growth is itself growing exponentially.” As Kurzweil reasons, “The implications include the merger of biological and nonbiological intelligence, immortal software-based humans, and ultra-high levels of intelligence that expand outward in the universe at the speed of light.”[4] This is not as far-fetched as it might sound. On the contrary, this statement may well be conservative. No one may be able to grasp what life will look like 50 years from now. Our limited perception cannot really imagine what technology will be capable of in 2070, and therefore we do not have the concepts and words to describe it. It would be like asking humans from the stone age to imagine cars (let alone self-driving cars) and making video calls over the internet (the inter-what?). With exponential growth, nothing much seems to happen for a while, and then suddenly everything seems to happen all at once. It is not possible to address compounding issues retroactively; by the time they are visible, it’s already too late. That’s why it’s strategic to err on the side of caution. This dynamic has occurred throughout human history and is not unique to the digital age. However, we just happen to be the generations alive to witness the inflection point where change moves beyond comprehension, but with minds and bodies the same as our ancestors. As Edward O. Wilson describes, “The real problem of humanity is the following: we have Paleolithic emotions, medieval institutions and godlike technology. And it is terrifically dangerous, and it is now approaching a point of crisis overall.”[5] Given that AGI is likely to exceed human intelligence and will have the ability to improve itself, the most honest prediction that experts can make is that they don’t know what will happen. It’s being built anyway. Ancient Human Nature

#### Err aff – Moore’s Law, quantum computing, and expert predictions make AGI inevitable

**Dilmegani 25** [Cem Dilmegani, computer engineer and Principal Analyst at AIMultiple with a degree in computer engineering from Bogazici University and a MBA from Columbia Business School, 2-18-2025, "When Will AGI/Singularity Happen? ~8,600 Predictions Analyzed," AIMultiple, https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/]/Kankee

We analyzed 8,590 scientists’, leading entrepreneurs’ and community’s predictions for quick answers on Artificial General Intelligence (AGI) / singularity timeline: Will AGI/singularity ever happen: According to most AI experts, yes. When will the singularity/AGI happen: Current surveys of AI researchers are predicting AGI around 2040.However, just a few years **before** the **rapid advancements** in large language models (LLMs), **scientists** were **predicting it around 2060**. Entrepreneurs are even more bullish, predicting it around ~2030. What is our current status: Although narrow AI surpasses humans inspecific tasks, a generally intelligent machine doesn’t exist, though someresearchers believe large language models demonstrate emerging generalistcapabilities.1 According to our AGI benchmark, machines are far from generatingeconomic value autonomously.Explore key predictions on AGI from experts like Sam Altman andDemis Hassabis, insights from five major AI surveys on AGI timelines, andarguments for and against the feasibility of AGI:Artificial General Intelligence timelineThis timeline outlines the anticipated year of the singularity,based on insights gathered from 15 surveys, including responses from 8,590 AI researchers, scientists, and participants in prediction markets:As you can see above, survey respondents are starting to think that singularity will take place earlier than expected.Understand what singularity is & why we fear itArtificial intelligence scares and intrigues us. Almost everyweek, there’s a new AI scare on the news like developers afraid of what they’vecreated or shutting down bots because they got too intelligent.2Most of these myths result from research misinterpreted by thoseoutside the AI and GenAI fields. Some stoke fear about AI because they mayprofit from more regulation or it may bring them more attention.The greatest fear about AI is singularity (also called ArtificialGeneral Intelligence or AGI), a system that combines human-level thinking withrapidly accessible near-perfect memory. According to some experts, singularityalso implies machine consciousness.Such a machine could self-improve and surpass human capabilities.Even before artificial intelligence was a computer science research topic,science fiction writers like Asimov were concerned about this. They weredevising mechanisms (i.e. Asimov’s Laws of Robotics) to ensure the benevolenceof intelligent machines which is more commonly called alignment research today.Results of major surveys of AI researchersWe examined the results of 10 surveys involving over 5,288 AIresearchers and experts, where they estimated when AGI/singularity might occur.While predictions vary, most surveys indicate a 50% probability ofachieving AGI between 2040 and 2061, with some estimating thatsuperintelligence could follow within a few decades.Expert Survey on Progress AI in 2023In October 2023, AI Impacts surveyed 2,778 AI researchers on whenAGI might be achieved. This survey included nearly identical question with the2022 survey. Based on the results, the high-level machine intelligence isestimated to occur until 2040.3Expert Survey on Progress in AI in 2022The survey was conducted with 738 experts who published at the2021 NIPS and ICML conferences, AI experts estimate that there’s a 50% chancethat high-level machine intelligence will occur until 2059.4Survey on what do ML researchers think about AI in 2022Katja Grace, co-founder of AI Impacts research project, conducteda survey with 356 AI experts. These experts were asked when they believe thereis a 50% probability that human-level AI will be achieved, which is defined asautonomous machines capable of performing all tasks more efficiently andcost-effectively than human workers.Half of the experts predicted this would happen before 2061, while90% estimated it would occur within the next century.5Survey of ML researchers’ forecasts on AI progress in 2019Baobao Zhang conducted a survey of 296 AI experts, asking them topredict when machines would surpass the median human worker in performing over90% of economically relevant tasks. Half of the respondents estimated thiswould happen before 2060.6AI experts survey on AGI timing in 2019The predictions of 32 AI experts on AGI timing7 are:45% of respondents predict a date before 2060.34% of all participants predicted a date after 2060.21% of participants predicted that singularity will never occur.Survey on AI’s potential impact of labor displacement in 2018Ross Gruetzemacher surveyed 165 AI experts to assess the potentialimpact of AI on labor displacement. The experts were asked to estimate when AIsystems would be capable of performing 99% of tasks for which humans arecurrently paid, at a level equal to or exceeding that of an average human.Half of the respondents predicted this milestone would be reachedbefore 2068, while 75% anticipated it would occur within the next 100 years.8AI experts in 2015 NIPS and ICML conferences survey in 2017In 2017 May, 352 AI experts who published at the 2015 NIPS andICML conferences were surveyed.9Based on survey results, experts estimate that there’s a 50%chance that AGI will occur until 2060. That said, there’s a significantdifference of opinion based on geography: Asian respondents expect AGI in 30 years,North Americans expect it in 74 years.Some significant job functions that are expected to be automateduntil 2030 are call center reps, truck driving, and retail sales.Future Progress in Artificial Intelligence survey in 2012/2013Vincent C. Muller, the president of the European Association forCognitive Systems, and Nick Bostrom from the University of Oxford, whopublished over 200 articles on superintelligence and artificial generalintelligence (AGI), conducted a survey of AI researchers. 550 participantsanswered the question: When is AGI likely to happen?10According to the results:The surveyed AI experts estimate that AGI will probably (over 50%chance) emerge between 2040 and 2050 and is very likely (90% chance) to appearby 2075.Once AGI is reached, most experts believe it will progress tosuper-intelligence relatively quickly, with a timeframe ranging from as littleas 2 years (unlikely, 10% probability) to about 30 years (high probability,75%).2009 survey with AI experts participating the in AGI-09 conferenceBased on the results of the survey with 21 AI expertsparticipating the in AGI-09 conference, it is believed that AGI will occuraround 2050, and plausibly sooner.11 You can see below their estimatesregarding specific AI achievements: passing the Turing test, passing thirdgrade, accomplishing Nobel worthy scientific breakthroughs and achievingsuperhuman intelligence.Figure 1: Results from the survey distributed to attendees of theArtificial General Intelligence 2009 (AGI-09) conferenceMicrosoft’s report on early experiments with GPT-4Microsoft Research studied an early version of OpenAI’s GPT-4 in2023. The report claimed that it showed greater general intelligence thanprevious AI models, performed at a human level in areas like math, coding, andlaw. This sparked debate on whether GPT-4 was a preliminary form of artificialgeneral intelligence. 12Community insightsWe also evaluated Metaculus community predictions on AGI whichinvolved the predictions of more than 3,290 participants:In 2022, 172 participants answered the question “When will an AIfirst pass a long, informed, adversarial Turing test?” and their prediction was2029.13In 2022, 81 participants answered the question “When will topforecasters expect the first Artificial General Intelligence to be developedand demonstrated?” and their prediction was 2035.14In 2020, 1,563 participants answered the question “When will thefirst weakly general AI system be devised, tested, and publicly announced?” andtheir prediction was 2026.15In 2020, 1,474 participants answered the question “When will thefirst general AI system be devised, tested, and publicly announced?” and theirprediction was 2030.16Insights from AI entrepreneurs & individual researchersAI entrepreneurs are also making estimates on when we will reachsingularity and they are more optimistic than researchers. This is expected asthey benefit from increased interest in AI.Here are the predictions of 12 of the most prominent AI entrepreneurs and researchers:Elon Musk expects development of an artificial intelligencesmarter than the smartest of humans by 2026.17Dario Amodei, CEO of Anthropic, expects singularity by 2026.18In February 2025, entrepreneur and investor Masayoshi Son predicted it in 2-3 years, (i.e. 2027 or 2028)19In March 2024, Nvidia CEO Jensen Huang predicted that within fiveyears, AI would match or surpass human performance on any test: 2029.20Louis Rosenberg, computer scientist, entrepreneur, and writer, by 2030.Ray Kurzweil, computer scientist, entrepreneur, and writer of 5national best sellers including The Singularity Is Near: Previously 2045,21 ,in 2024, 2032.22 In 2023, Hinton believed that it could take 5-20 years.23Demis Hassabis, founder of DeepMind, by 2035.24Sam Altman, CEO of OpenAI, by 2035. He mentioned “a few thousanddays” in 2024 in his blog “The Intelligence Age”.Ajeya Cotra, an AI researcher, analyzed the growth of trainingcomputation and estimated a 50% chance that AI with human-like capabilities will emerge by 2040.25Patrick Winston, MIT professor and director of the MIT ArtificialIntelligence Laboratory from 1972 to 1997: He mentioned 2040 while stressingthat while it would take place, it is a very hard-to-estimate date.Jürgen Schmidhuber, co-founder at AI company NNAISENSE anddirector of the Swiss AI lab IDSIA, by 2050. 26Learning from past over-optimism in AI predictionsKeep in mind that AI researchers were over-optimistic before.Examples include:AI pioneer Herbert A. Simon in 1965: “machines will be capable,within twenty years, of doing any work a man can do.”27Japan’s Fifth Generation Computer in 1980 had a ten-year timelinewith goals like “carrying on casual conversations”.28This historical experience contributed to most current scientistsshying away from predicting AGI in bold time frames like 10-20 years, but, thishas changed with the rise of generative AI. Why experts believe AGI is inevitable: Key arguments and evidence Reaching AGI may seem like a wild prediction, but it seems like quite a reasonable goal when you consider these facts: Human intelligence is fixed unless we somehow merge our cognitive capabilities with machines. Elon Musk’s neural lace startup aims to do this but research on brain-computer interfaces is in the early stages.29 Machine intelligence depends on algorithms, processing power, and memory. Processing power and memory have been growing at an exponential rate. As for algorithms, until now we have been good at supplying machines with the necessary algorithms to use their processing power and memory effectively. Considering that our intelligence is fixed and machine intelligence is growing, it is only a matter of time before machines surpass us unless there’s some hard limit to their intelligence. We haven’t encountered such a limit yet. Below is a good analogy for understanding exponential growth. While machines can seem not highly intelligent right now, they can grow quite smart, quite soon. If classic computing slows its growth, quantum computing could complement it Classic computing has taken us quite far. AI algorithms on classical computers can exceed human performance in specific tasks like playing chess or Go. For example, AlphaGo Zero beat AlphaGo by 100-0. AlphaGo had beaten the best players on earth.30 However, we are approaching the limits of how fast classical computers can be. Moore’s law, which is based on the observation that the number of transistors in a dense integrated circuit double about every two years, implies that the cost of computing halves approximately every 2 years. On the other hand, most experts believe that Moore’s law is coming to an end during this decade.31 However, there are efforts to keep improving efficiency of compute. For example, DeepSeek surprised global markets with its R1 model by delivering a reasoning model at a fraction of the cost of its competitors like OpenAI. Quantum Computing, which is still an emerging technology, can contribute to reducing computing costs after Moore’s law comes to an end. Quantum Computing is based on the evaluation of different states at the same time whereas classical computers can calculate one state at one time. The unique nature of quantum computing can be used to efficiently train neural networks, currently the most popular AI architecture in commercial applications. AI algorithms running on stable quantumcomputers have a chance to unlock singularity. For more information, read quantum computing. Why some experts believe that we will not reach AGI?

#### AGI is on the brink – tech current exists for AGI creation tomorrow

**Tivy 23** [Wolf Tivy, Editor-At-Large of Palladium Magazine and former junior engineer at ZincNyx Energy Solutions with a BS in mechanical engineering from the British Columbia Institute of Technology, 8-10-2023, “Artificial General Intelligence Is Possible and Deadly”, Palladium Magazine, https://www.palladiummag.com/2023/08/10/artificial-general-intelligence-is-possible-and-deadly/]/Kankee

The real problem of AGI feasibility is philosophical: restating McCarthy’s conjecture, can we untangle what we mean by intelligence enough to build an algorithm for it? Judging by the practical and cross-cultural robustness of the concept, we probably can. We just don’t know how yet. Modern Engineering Can Build AGI In a sunken ship from ancient Greece, divers dredged up a complex assembly of corroded gears and wheels that computed the positions of the planets. Dubbed the Antikythera mechanism, this artifact was evidence not just of one skilled craftsman but of a whole science and proto-industry advanced beyond anything previously understood. In particular, it was evidence of some understanding of the idea of mechanical calculation. Archimedes himself, or some other scientist of the time, may have dreamed of using these principles of clockwork computation to build a mind. Perhaps they would have been inspired by the god Hephaestus’s golden robotic serving girls from the Iliad. But if they did dream of it, their engineering wasn’t ready. They didn’t have powerful enough computers. Furthermore, they probably did not have the calculus, information theory, probability theory, computer science, and the other still-unknown parts of mathematics and philosophy necessary to even design such a thing. There may likewise be some crucial faculty we are missing that would prevent us from building AGI. But we have many things the ancients didn’t have, which suggests that we are very close in most of the key engineering capacities. Engineers in the twenty-first century have built computer-controlled robots that walk around, dance, talk, hear and follow commands, and see the world around them. Many factories have significant robotic automation already, limited mostly by the lack of intelligence of the machines. We have a basic understanding of how the brain’s neurons act as computational elements, how much computation they do, and how many of them there are. By some estimates, we are now crossing the threshold where a human brain’s worth of computation is available off the shelf. We have a whole industry of people who can build computer programs to implement whatever algorithmic principle we can discover. The major missing piece is the key ideas of artificial general intelligence themselves. Given those ideas, the engineering side seems poised to put them into action. If we grant the AI field’s major conjecture—that general intelligence can be specified to the point that practical algorithms can be written to implement it—there are several different paradigms of how to even do this. “Good old fashioned AI” (GOFAI) attempts to build a system that symbolically reasons from evidence, assumptions, knowledge, heuristics, and goals to plans and actions. “Connectionist” AI, based on neural nets, dispenses with symbolic abstraction and deals directly with complex but low-level black-box calculations of perception and reflexive response. We can also imagine a “genetic” approach that evolves a complex ecosystem of interacting sub-processes selected by some overall logic. These are just some of the approaches conceived so far, and there may well be others. Within each individual paradigm, there are a dozen or more different parameters. What kinds of assumptions and heuristics? What kind of logic system? How precise versus approximate? What kind of neural architectures? How are they trained? Is training distinct from “inference?” What kind of subprocesses? Do we need a combination of these approaches? And so on it goes. We have a wealth of threads to pull on, and any one of them could yield the key insights. The recent wave of progress in deep learning resulted from the unexpected effectiveness of applying GPU acceleration to back-propagation-based training ideas invented in the late 1980s. In between then, neural nets had mostly stagnated as an approach. Where deep learning goes next, and if it goes anywhere novel at all, is hard to know. The next major breakthrough could be another deep learning architecture like the “attention”-based transformer, but it could also come from somewhere else entirely. Perhaps some breakthrough in proof theory or symbol learning could suddenly make GOFAI viable the way deep learning suddenly made neural nets viable. Or the field could stagnate for another 20 years. Further progress may depend on some new branch of mathematics developed by an unrelated neo-Pythagorean cult. The whole thing may even depend on new philosophy or theology. It may be that no one currently working on these problems has the right mix of obscure skills and knowledge. However, it may also be that we are very close, and the right person working in their basement could publish a working AGI architecture tomorrow. Many of the details of AGI are unknown unknowns, which are impossible to predict. Discussions about “AI timelines” are therefore not well-grounded. But we can say with significant confidence that AGI is possible and that twenty-first-century engineering is ready to put it into practice if the scientific and philosophical breakthroughs come through.

## Negative

### AT: AI Arms Races

#### China loses – failing semiconductor manufacturing, low prioritization, and other challenges

**Behrens 23** [Eva Behrens, AI Policy Researcher and Masters in Innovation, Public Policy, and Public Value at University College London, 04-11-2023, “The AGI Race Between the US and China Doesn’t Exist”, Eva’s Substack, https://evabehrens.substack.com/p/the-agi-race-between-the-us-and-china]/Kankee

The AGI Race Between the US and China Doesn’t Exist. When I write “China”, I refer to the political and economic entity, the People’s Republic of China, founded in 1949. Leading US AI companies are currently rushing towards developing artificial general intelligence (AGI) by building and training increasingly powerful large language models (LLMs), as well as building architecture on top of them. This is frequently framed in terms of a great power competition for technological supremacy between the US and China. However, China has **neither** the resources nor **any** interest in competing with the US in developing artificial general intelligence (AGI) primarily via scaling Large Language Models (LLMs). In brief, China does not compete with the US in developing AGI via LLMs because of: Resources and technology: China does not have access to the computational resources[1] (compute, here specifically data centre-grade GPUs) needed for large-scale training runs of large language models. This gap will only widen over time; China is failing to develop a domestic semiconductor industry, despite massive efforts to do so, and is increasingly **cut** **off** from international semiconductor supply chains. Political goals: The Chinese Communist Party (CCP) is above all concerned with ensuring social stability and staying in power. It will not tolerate large language models that cannot reliably be controlled and prevented from leaking sensitive information, raising politically sensitive topics, or challenging the regime. Reliably prohibiting LLMs from doing any of these things is an unsolved technical problem. This has and will continue to stifle the development and deployment of Chinese LLMs for the Chinese market, as a single political faux pas will spell **doom** for the companies involved. Therefore, as it currently stands, there is no AGI race between China and the US. China Cannot Compete with the US to Reach AGI First Via Scaling LLMs Today, training powerful LLMs like GPT-4 requires massive GPU clusters. And training LLMs on a larger number of GPUs is the most common and reliable way of increasing their capabilities. This is known as the bitter lesson in machine learning; increasing the compute a model is trained on leads to higher progress in the model's capabilities than any changes in the algorithm's architecture. So if China wants to “win the AGI race” against the US, it would need to maximise its domestic GPU capacities. However, China lacks domestic semiconductor production facilities for producing data centre-grade GPUs. So to increase domestic GPU capacities, it has to import GPUs or semiconductor manufacturing equipment (SME). However, since the US introduced its export controls on advanced chips and SMEs in October 2022, China can no longer purchase GPUs or SMEs abroad. And with the semiconductor industry advancing at lightning speed, China is falling further and further behind every day2. These policies are just the most recent ones in a row of ongoing measures by the US restricting the transfer of semiconductor technology to Chinese companies, beginning with Huawei in 2019. In early 2023, the Netherlands and Japan, which play a key role in the global semiconductor manufacturing chain, joined the US and introduced SME export controls targeting China. Crucially, these combined export controls apply to ASML’s most advanced lithography machines, the only machines in the world which can produce cutting-edge chips. The problem for China lies not so much in designing cutting-edge GPUs, but in designing and building SMEs to produce the advanced GPUs needed for training LLMs, a far more challenging task. But China has been trying - and failing - to build domestic semiconductor manufacturing chains for cutting-edge chips since at least the 1990s. In the 1990s, the state-sponsored Projects 908 and 909 only resulted in one company that today produces chips using 55nm process technology (technology from the mid-2000s, producing less complex chips used for applications like the Internet of Things and electric vehicles) and occupies around 3% of global foundry market share. In its 13th Five-Year Plan (2016-2020), the CCP listed building a domestic semiconductor industry as a priority, and in the 14th Five-Year Plan (2021-2025), developing domestic semiconductor supply chains appears as one of the country’s main goals in industrial and economic policy. Since then, China has increased domestic chip production capacities, but still only succeeds at producing larger, less complex chips. And achieving even this success was possible largely due to the forced transfer of intellectual property from US-American and Taiwanese semiconductor firms. This is now much harder with strict export controls from the US on advanced GPUs and a broader plethora of measures aimed at restricting the transfer of chips and SME technology and design to Chinese firms. In addition, in the summer of 2022, the Chinese government started investigating multiple senior figures who were responsible for industrial policy and state investment programmes to create a domestic cutting-edge chip industry starting in 2014. These investigations are based on “opaque allegations” of officials breaking the law. In early 2023, China then changed its strategy from granting generous subsidies to semiconductor manufacturing companies to focusing on a small number of potential national champions. This recent policy shift indicates that the Chinese leadership has decided that prior strategies have failed. In addition, the new policy programme has a smaller budget, which points to the Chinese government prioritising other challenges (more on this below). On a general note, Chinese industrial policy officially has the goal of building industry and increasing economic growth, but its actual goal is first and foremost to preserve CCP power and social stability. Chinese industrial policy routinely fails at its stated goals precisely because it is optimised for preserving CCP power, which it has succeeded at since the late 1970s. This goes beyond the semiconductor industry and e.g. also causes failures in China’s green transition. For these reasons, the probability that the Chinese government will succeed at or even properly try to build a domestic semiconductor industry capable of producing cutting-edge, data centre-grade GPUs is extremely low. On top of everything, China is facing food security, public health, and demographic challenges, and the rising cost of Chinese labour is threatening its economic model as the world's workbench. These combined challenges pose a more significant threat to the CCP’s hold on power than lagging behind the USA in AGI, so the CCP will most likely prioritise these challenges and continue to reduce funds available for building domestic GPU supply chains. And the challenges China faces today will only get more serious, as its population ages further and climate change increases the frequency of extreme weather events. Hence, in combination with the growing lead of the US in semiconductor technology, as time passes, it will become even more difficult for China to catch up to the US than it is today. For Political Reasons, China Won’t Compete with the US to Reach AGI First Via Scaling LLMs

#### China loses – authoritarianism stifles AGI development

**Behrens 23** [Eva Behrens, AI Policy Researcher and Masters in Innovation, Public Policy, and Public Value at University College London, 04-11-2023, “The AGI Race Between the US and China Doesn’t Exist”, Eva’s Substack, https://evabehrens.substack.com/p/the-agi-race-between-the-us-and-china]/Kankee

For Political Reasons, China Won’t Compete with the US to Reach AGI First Via Scaling LLMs The CCP is not interested in reaching AGI by scaling LLMs. Reliably controlling powerful LLMs is an unsolved technical problem 3: for instance, it is currently impossible to get an LLM to reliably never output specific words or phrases or touch on certain topics. Given this, any sufficiently powerful Chinese LLM is guaranteed to make a politically taboo statement sooner or later, like referencing the Tian’anmen Square massacre. This makes this line of AI development too politically sensitive and ultimately undesirable in the Chinese environment. Instead, the CCP wants and needs reliable, narrow, controllable AI systems to censor internet discourse, improve China’s social credit system, and expand citizen surveillance. This is reflected in the topics Chinese and US AI researchers write about. Chinese researchers are overrepresented among highly cited papers on AI applications in surveillance and computer vision, while US researchers publish more leading papers on deep learning. Considering this, one of two things will happen: Either Chinese LLMs will continue to be weaker, due to a lack of compute or by design to comply with government guidelines, or a Chinese developer will release a powerful LLM on the Chinese internet, it will cause a political scandal by not conforming with censorship rules or state-issued guidelines, and the government will shut it down. Chinese companies will only build LLMs that reliably comply with Chinese censorship laws and upcoming laws or guidelines on generative AI. To ensure models never go off-script, companies will have to make them weaker, and very carefully control who gains access to the models or the APIs. For example, the March 2023 demo for Baidu’s new generative language model Ernie Bot consisted only of pre-recorded video clips, disappointing the public, and so far, only businesses can apply for licenses to use the model. Baidu advertised Ernie Bot as a multimodal model, meaning it should be able to generate responses to different forms of input, e.g. including text and images. However, Ernie Bot lacks the ability to interpret images and generate a text response. Furthermore, users who had the opportunity to test Ernie Bot report that it performed slightly worse than GPT-3, a model that is now almost three years old, and, unlike ChatGPT, shirks away from answering questions on political topics. For context, these reviews were published in late March 2023, when OpenAI had already released the more powerful GPT-4. One reviewer summarised the situation by commenting on the model’s “bearable mediocrity” and writing it is “probably good enough for the Chinese market”. In short, Ernie Bot is just good enough to probably for now satisfy the demand in a market where all serious competitors are banned. In direct competition with the generative language models from companies like OpenAI, Ernie Bot would not be able to compete. If a Chinese developer ever releases an LLM that is so powerful it inevitably oversteps censorship rules at some point, the Chinese government will block it and crack down on the company that released it. In the past, the CCP has held tech companies accountable for any undesirable content posted by users of their platforms, and according to the new draft guidelines on generative AI, it will do the same regarding LLMs. The temporary disappearance of Jack Ma in 2020 when the CCP decided that his company Alibaba had become too powerful is another cautionary tale for Chinese tech CEOs to not challenge the CCP. Indeed, the new drafts on guidelines for generative AI in China from April 20234, published shortly after I first wrote this piece, give an idea of the intentions of the CCP regarding the regulation of generative AI models, including LLMs. Specifically, these draft guidelines state that “the content generated by generative artificial intelligence [...] must not contain subversion of state power”. This statement is a clear indicator of the CCP’s intention to prohibit any generative AI model from being deployed that may at some point produce politically undesirable outputs. But So under the draft guidelines, powerful LLMs would be completely prohibited. No Global AGI Race in Sight In conclusion, China is not in a position to build AGI via scaling LLMs because it lacks the resources, technology, and political will to do so. First, China doesn’t have the necessary computing resources, data centre-grade GPUs, to continue doing increasingly large LLM training runs. China is unable to purchase these GPUs, primarily designed in the US and manufactured outside of China, because the US and its allies have placed strict export controls on this technology, as well as the SMEs needed to produce sufficiently powerful GPUs. Second, China is also unable to manufacture data centre-grade GPUs domestically. This situation is only worsening as the CCP will have to deprioritise industrial policy programmes to develop domestic cutting-edge SMEs and GPUs in favour of dealing with more pressing challenges, such as demographic change and the rising cost of Chinese labour. Lastly, the Chinese government has no interest in reaching AGI via powerful LLMs, since LLMs cannot be controlled reliably. The Chinese government made this clear in its recent draft guidelines on generative AI. Instead, the CCP is interested in using narrow, controllable AI systems for surveillance. As a result, AI companies in China will either continue to release weak LLMs, like Baidu’s Ernie Bot or will be cracked down on if they release a sufficiently powerful and therefore uncontrollable model. In short, there is no AGI race between the USA and China. Hence, framing the development of AGI as a competition between the USA and China, in which the US must compete to win, is misleading. All companies with leading AGI ambitions are located in the US and do not have direct competitors in China. If concerns over a US-China AGI race are the reason why your company is accelerating its progress towards AGI, then your concerns are unfounded and you should stop.

### AT: ChatGPT = AGI

#### AGI is categorically different from narrow AI – AGI is currently impossible with existing tech and know-how

**Marr 24** [Bernard Marr, world-renowned futurist, board advisor and author of [Generative AI in Practice: 100+ Amazing Ways Generative Artificial Intelligence is Changing Business and Society](https://protect-us.mimecast.com/s/02BlCG6wq3fmpMVpsKwZx8?domain=amazon.co.uk) and written over 20 best-selling and award-winning books and advises and coaches many of the world’s best-known organisations, 5-20-2024, “The Crucial Difference Between AI And AGI”, Forbes, https://www.forbes.com/sites/bernardmarr/2024/05/20/the-crucial-difference-between-ai-and-agi/]/Kankee

The Theoretical Landscape Of AGI In stark contrast to the specific applications of current AI systems, AGI represents a theoretical pinnacle of this technology. Unlike specialized AI, AGI would be capable of understanding and reasoning across a broad range of tasks. It would not only replicate or predict human behavior but also embody the ability to learn and reason across diverse scenarios, from creative endeavors to complex problem-solving. To do that, it would require not just Intelligence but also emotional and contextual awareness. This type of Intelligence could potentially manage diverse and complex tasks that require creativity, emotional Intelligence, and multi-dimensional thinking—capabilities far beyond the reach of today's AI. However, the journey toward AGI is hindered by our current understanding and technological limitations. Building machines that truly understand and interact with the world like humans involves not just technical advancements in how machines learn, but also profound insights into the nature of human Intelligence itself. Current AI lacks the ability to fully comprehend context or develop a worldly understanding, which is critical for tasks that humans navigate seamlessly. As AI technology progresses, grasping the profound distinctions between AI and AGI is essential. While AI already improves our daily lives and workflows through automation and optimization, the emergence of AGI would be a transformative leap, radically expanding the capabilities of machines and redefining what it means to be human.

#### Large language models cannot be AGI

**Emmert-Streib 24** [Frank Emmert-Streib, Predictive Society and Data Analytics Lab Department, Tampere University, 5-7-2024, “Is ChatGPT the way toward artificial general intelligence”, Discover Artificial Intelligence, , https://link.springer.com/article/10.1007/s44163-024-00126-3]/Kankee

4 Shortcomings of ChatGPT for AGI For the following discussion, we assume that ChatGPT takes the position of the agent in Fig. [1](https://link.springer.com/article/10.1007/s44163-024-00126-3#Fig1). Based on this configuration, we are now able to identify three important shortcomings of the current version of ChatGPT preventing it from reaching AGI. 1. Currently, ChatGPT cannot interact with the environment via an action. 2. Currently, ChatGPT does not have the inner structure (policy function) of an agent. 3. The environment consists only of text. We would like to remark that the first two limitations are related because an agent without policy function cannot select an action but even with a policy function one could disallow the execution of an action which would prevent the interaction with an environment. While ChatGPT generates an output, this output cannot directly affect the environment, making it fundamentally different from the actions of an agent. The third shortcoming is related to the fact that, in contrast to a robot, ChatGPT is a natural language processing system that reads text and generates text. That means ChatGPT lives per se in a virtual environment. This removes, in a natural way, the embodiment problem from robotics discussed above without the need for any approximation. However, this is a severe simplification and entirely different to the situation as faced by humans, animals or robots. We would also like to highlight that the text-based nature of the environment is a property of the environment rather than a technical limitation. As such (3) is different to (1) and (2) which can be changed by a modification of ChatGPT. This indicates that, as it stands, ChatGPT fails to meet the fundamental criteria of an agent necessary for achieving AGI. However, this does not diminish its ability to amaze its users. Nonetheless, it’s important to recognize that ChatGPT, in its current state, is not and cannot evolve into an AGI. Therefore, in the subsequent sections, we explore potential extensions of ChatGPT to gain insights into the most realistically achievable objectives. 5 GPT-in-the-loop as artificial special intelligence For the following discussion, we will accept shortcoming (3) because it reflects the situation as is (corresponding to a property of the environment) but improve upon shortcomings (1) and (2). This will lead to the definition of artificial special intelligence (ASI). While the goal of AGI is the development of autonomous systems that possess the cognitive capabilities and general intelligence similar to that of human beings one can limit this ambitious goal. Specifically, let’s content ourselves with tasks only involving text. That means in this case, the environment consists only of text and no images, audio signals etc can occur. Hence, in this case the environment would be a text world. Assuming the environment is a text world, this has two important implications. 1. An agent interacting with such an environment does not need physical sensors nor actors because the text world is virtual. 2. An agent living in a text world does not require embodiment because no “physical” entities are involved in this perception-action cycle. These two implications enable an agent with more limited capabilities due to a simpler environment than required for AGI. For this reason we call this restricted case artificial special intelligence (ASI) because an optimal rational agent lives in a text world. Definition 1 (Artificial special intelligence [ASI]) Artificial special intelligence (ASI) is an agent’s capability to perform any intellectual task that a human being can, based on text data. Regarding the chosen name, we would like to remark that we use artificial special intelligence (ASI) to mirror the same sentiment as in artificial general intelligence (AGI) but emphasizing a particular or “special” situation. More precisely it could be called artificial text intelligence (ATI), however, if considered in isolation, this asymmetry between AGI and ATI could be overlooked. For this reason, we prefer the name ASI over ATI, as it maintains consistency with the sentiment of AGI and underscores the specificity of an application. Next, we address the need for an agent to act on the environment. Since ChatGPT is currently not an agent as in Fig. [1](https://link.springer.com/article/10.1007/s44163-024-00126-3#Fig1), we need to extend its structure to convert it into such an agent. That means, we need to place ChatGPT in-the-loop (ITL) and for this reason we call the resulting system GPT-ITL. Specifically, this would require the following modifications. 1. GPT-ITL would need access to the environment to collect new input (perception). 2. GPT-ITL would need to receive a reward from the environment. 3. The inner structure of GPT-ITL would need to be adjusted by including a policy. 4. GPT-ITL would need access to the environment by writing its output into the text world (action). At the moment, all steps seem feasible from a technical point but ethically there are issues, especially with the last step. The problem with the last step is that it would give autonomy to the system to publish its output publicly. However, this would allow that GPT-ITL could spread misinformation because false output/answers could enter the public domain, e.g., on social media platforms. Even more severely, GPT-ITL could manipulate social media users and even persuade them to commit certain doings including crimes. Looking from a fundamental perspective onto GPT-ITL, it is clear that due to our assumption of a simplified environment consisting of a text-world, GPT-ITL could at best reach an ASI. Hence, even GPT-ITL would not be capable of reaching AGI. 6 Modified version of ChatGPT with private input: LLM-PI Now we turn from fundamental considerations to practical ones by outlining a way forward. For the moment, this approach will avoid ethical concerns encountered for GPT-ITL, instead, this will be considered in Sect. [7](https://link.springer.com/article/10.1007/s44163-024-00126-3#Sec7). For this approach, we start again from ChatGPT and augment it with a private input (PI). We call this model, LLM-PI where LLM (large language model) indicates that it could be GPT or any other LLM. Figure [2](https://link.springer.com/article/10.1007/s44163-024-00126-3#Fig2) outlines the main components of LLM-PI. Importantly, this system is not allowed to act on the environment (violet box and its connecting lines are not part of this system; see below). That means the perception-action loop is open and, hence, the resulting agent, corresponding to LLM-PI, is no longer autonomous. Aside from this, LLM-PI has a separate private input (PI) channel allowing the user to supply text data, e.g., corresponding to documents, notes or files that are not available from the environment. In this way, problem specific or confidential information can be supplied that allows LLM-PI the answering of questions requiring either private or specific information that is not part of the public text world (environment). We would like to emphasize that the crucial difference between LLM-PI and the current form of ChatGPT is the presence of the private input channel. While this may appear straightforward technically, it entails more than just a static input. In order to make this additional information actionable, the LLM must undergo another round of fine-tuning for learning from the provided input. While theoretically feasible, this process is time and resource-intensive in practice. Consequently, users may experience delays as they await the completion of this fine-tuning process before receiving answers to their questions, potentially rendering real-time conversations impractical at present. In addition to the computational overhead, implementing a private input channel offers a distinct advantage for the personalization of a LLM. Over time, this feature would enable the system to tailor responses to individual users, providing more authentic interactions that align with personal preferences. However, this personalized approach raises ethical concerns surrounding user privacy, as safeguards would be necessary to prevent potential abuse by third parties. Nonetheless, these concerns are manageable and can be addressed through appropriate policies, representing a different set of challenges compared to those encountered with AGI. 7 Gating the loop for LLM-PI: gLLM-PI The last extension we want to discuss provides an approach for the implementation of such a policy. Specifically, we suggest a gating mechanism for actions. Such a gating mechanism for actions allows to close the loop to the environment but in a gated manner. Hence, the gating would allow a moderation of actions in order to control ethical implications potentially incurred by the actions. While this controls the autonomy of the agent it mediates ethical concerns. With regard to the LLM-PI discussed in the previous section, one could start with such an optimized model allowing private input but augment it by an additional functionality enabled via the closed but gated perception-action-cycle. Overall, this brings us back to similar capabilities as discussed earlier in Sect. [5](https://link.springer.com/article/10.1007/s44163-024-00126-3#Sec5) and shows how to approach an artificial special intelligence (ASI) system that is limited to a text environment. To distinguish this model from others, we call it gLLM-PI (gated large language model with private information). In Fig. [2](https://link.springer.com/article/10.1007/s44163-024-00126-3#Fig2), we show gLLM-PI that is obtained by adding gated actions (violet box) and its connecting lines to LLM-PI. Importantly, by using different gating mechanisms for selectively permitting actions different forms of ASI are enabled. Potentially, this allows us to further sub-categorize ASI into more refined classes that represent different types or severities of ethical concerns. In order to get a better understanding about this, let’s consider a few examples. Example 1 The following list gives hierarchical restrictions applied to a ASI via gated actions. We start at the top of such an ASI corresponding to the most liberal form and work our way down to the most restricted form. Gated actions: 1. Politeness and courtesy 2. 1 + political correctness 3. 2 + falsifiability 4. 3 + expert in medicine The first restriction (1) enforces a polite and courteous communication, thereby prohibiting the use of inappropriate language. This results in a formal rather than colloquial discussion. The second restriction (2) additionally (1 + means the first restriction and the second) requires that the communication be politically correct. Unfortunately, this restriction would be country- and time-specific and not universal, as the definition of political correctness varies widely among countries and changes over time. This is in contrast to restriction (3), which requires in addition (2 + means the second and the third) the falsifiability of statements, limiting communication to a factual basis with a clear distinction between facts and hypotheses. Finally, restriction (4) requires in addition that the communication is limited to the field of medicine. Overall, this results in a particular ASI where the gLLM-PI is similar to a medical expert. We would like to note that by replacing restriction (4) with, e.g., “3 + expert in economics” we obtain another ASI but now as an expert in economics. Example 2 Another instance of a ASI for other gated actions is given by the following restrictions. Gated actions: 1. Politeness and courtesy 2. 1 + falsifiability 3. 2 + expert in medicine While this looks very similar to the gated actions for Example [1](https://link.springer.com/article/10.1007/s44163-024-00126-3#FPar2), it does not contain restrictions about political correctness (see above). However, as discussed above, such restrictions are not only country-specific but change also over time as political legislation can change. That means the gLLM-PI for Example [2](https://link.springer.com/article/10.1007/s44163-024-00126-3#FPar3) would be less restricted than the gLLM-PI for Example [1](https://link.springer.com/article/10.1007/s44163-024-00126-3#FPar2). As a result, one could study the effect different gated actions have on the resulting communication and the obtained ASI. We would like to point out that our description of a gLLM-PI is unlike, e.g., standard agent types for ordinary artificial intelligence as described in [[33](https://link.springer.com/article/10.1007/s44163-024-00126-3#ref-CR33)]. The reason for this difference is due to our specific situation given by the text environment with which GPT-ITL can interact. In summary, these examples illustrate various versions of a gLLM-PI, including ethical forms that can be studied and cross-investigated to enhance our understanding of an ASI. 8 Discussion While there is currently a great deal of excitement surrounding ChatGPT, and it certainly has the potential to contribute significantly to research and education, the hope (or fear) that it will achieve AGI status is not justified. Instead, as discussed in this paper, several extensions are necessary. Even with these extensions, ChatGPT would at best achieve artificial special intelligence (ASI) status, primarily due to its limitation to a text-based environment. Still, the idea of studying such an ASI is intriguing because of the following reasons. 1. No embodiment problem: Usually, embodiment requires, e.g., a robot to be situated within a (real) environment to perceive and act correspondingly. However, this requires various approximations that all rely on assumptions. Instead, gLLM-PI lives in a space where only text exists. Hence, the difference between a physical world and a virtual word does not exist and no approximations to bridge both worlds are needed. Still, such a world is not a toy example but represents real texts as exchanged on all levels of society. 2. Categories of actions: The most problematic aspect of a closed loop of gLLM-PI would certainly be its possibility to performing actions. However, there are different categories of actions having different consequences. For instance, there is a difference if gLLM-PI would be allowed to upload a report to a public repository, or if it would be allowed to make an online purchase. While the former might lead to misinformation the latter triggers with certainty a direct action by a third party. That means by carefully categorizing actions of gLLM-PI one could moderate ethical implications and real-world consequences. 3. Declaration of gLLM-PI generated text: Another problems that could result from generated text by gLLM-PI is the spread of misinformation. This problem could be mitigated by the declaration of gLLM-PI generated text. Generally, it is always important to consider the source of information or text and a declaration would eliminate anonymity and promote source transparency. This would allow for a more informed assessment by readers of such a text. Technically, by using blockchain technology one could ensure that there is no tampering with such a declaration. In summary, these are just a few aspects and approaches that could be studied to allow for a safe transition from ChatGPT in its current form to an ASI via a number of intermediate steps provided by different versions of gLLM-PI. If this paves also the way to a genuine AGI remains to be seen because it requires substantial extensions that would lead us beyond a text environment. 9 Conclusion We showed in this paper that, for principle reasons, ChatGPT is neither an artificial general intelligence (AGI) nor can it naturally evolve into one. That means without significant modifications of ChatGPT this is not feasible. A main problem for this is the confinement into a "text world" that does not require or enable embodiment to interact via sensors and actors with the environment. However, by introducing the notion of an artificial special intelligence (ASI) that is limited to an environment consisting only of text, a modified version of ChatGPT, we called gLLM-PI (gated large language model with private information), could transition into an ASI. In fact, the limitation to a "text world" seems to open new avenues for the exploration of the perception-action cycle that is an eminent component for the designing of an optimal rational agent in any setting. Furthermore, we showed that the gating mechanism for actions that closes the PAC in a controlled manner allows the implementation of policies for preventing ethical concerns that are inherently present for AGI. Hence, the gLLM-PI model has not only the potential to enhance our understanding of intelligent systems but to do this by moderating ethical considerations.

#### OpenAI admits themselves in contracts that LLMs are not AGI – ChatGPT must produce 100 billion dollars to qualify as AGI

**Pillay 25** [Tharin Pillay, TIME Editorial Fellow and Master of Laws at University College London, 01-08-2025, “How OpenAI’s Sam Altman Is Thinking About AGI and Superintelligence in 2025”, TIME, https://time.com/7205596/sam-altman-superintelligence-agi/]/Kankee

In a recent interview with Bloomberg, Altman said he thinks “AGI will probably get developed during [Trump’s] term,” while noting his belief that AGI “has become a very sloppy term.” Competitors also think AGI is close: Elon Musk, a co-founder of OpenAI, who runs AI startup xAI, and Dario Amodei, CEO of Anthropic, have both said they think AI systems could outsmart humans by 2026. In the largest survey of AI researchers to date, which included over 2,700 participants, researchers collectively estimated there is a 10% chance that AI systems can outperform humans on most tasks by 2027, assuming science continues progressing without interruption. Others are more skeptical. Gary Marcus, a prominent AI commentator, disagrees with Altman that AGI is “basically a solved problem,” while Mustafa Suleyman, CEO of Microsoft AI, has said, regarding whether AGI can be achieved on today’s hardware,“the uncertainty around this is so high, that any categorical declarations just feel sort of ungrounded to me and over the top,” citing challenges in robotics as one cause for his skepticism. Microsoft and OpenAI, which have had a partnership since 2019, also have a financial definition of AGI. Microsoft is OpenAI’s exclusive cloud provider and largest backer, having invested over $13 billion in the company to date. The companies have an agreement that Microsoft will lose access to OpenAI’s models once AGI is achieved. Under this agreement, which has not been publicly disclosed, AGI is reportedly defined as being achieved when an AI system is capable of generating the maximum total profits to which its earliest investors are entitled: a figure that currently sits at $100 billion. Ultimately, however, the declaration of “sufficient AGI” remains at the “reasonable discretion” of OpenAI’s board, according to a report in The Information. At present, OpenAI is a long way from profitability. The company currently loses billions annually and it has reportedly projected that its annual losses could triple to $14 billion by 2026. It does not expect to turn its first profit until 2029, when it expects its annual revenue could reach $100 billion. Even the company’s latest plan, ChatGPT Pro, which costs $200 per month and gives users access to the company’s most advanced models, is losing money, Altman wrote in a post on X. Although Altman didn’t explicitly say why the company is losing money, running AI models is very cost intensive, requiring investments in data centers and electricity to provide the necessary computing power. Pursuit of superintelligence OpenAI has said that AGI “could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility.” But recent comments from Altman have been somewhat more subdued. “My guess is we will hit AGI sooner than most people in the world think and it will matter much less,” he said in December. “AGI can get built, the world mostly goes on in mostly the same way, things grow faster, but then there is a long continuation from what we call AGI to what we call superintelligence.” In his most recent post, Altman wrote, “We are beginning to turn our aim beyond [AGI], to superintelligence in the true sense of the word. We love our current products, but we are here for the glorious future.” He added that “superintelligent tools could massively accelerate scientific discovery and innovation well beyond what we are capable of doing on our own, and in turn massively increase abundance and prosperity.” This ability to accelerate scientific discovery is a key distinguishing factor between AGI and superintelligence, at least for Altman, who has previously written that “it is possible that we will have superintelligence in a few thousand days.” The concept of superintelligence was popularized by philosopher Nick Bostrom, who in 2014 wrote a best-selling book—Superintelligence: Paths, Dangers, Strategies—that Altman has called “the best thing [he’s] seen on the topic.” Bostrom defines superintelligence as “any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest”—like AGI, but more. “The first AGI will be just a point along a continuum of intelligence”, OpenAI said in a 2023 blog post. “A misaligned superintelligent AGI could cause grievous harm to the world; an autocratic regime with a decisive superintelligence lead could do that too.” These harms are inextricable from the idea of superintelligence, because experts do not currently know how to align these hypothetical systems with human values. Both AGI and superintelligent systems could cause harm, not necessarily due to malicious intent, but simply because humans are unable to adequately specify what they want the system to do. As professor Stuart Russell told TIME in 2024, the concern is that “what seem to be reasonable goals, such as fixing climate change, lead to catastrophic consequences, such as eliminating the human race as a way to fix climate change.” In his 2015 essay, Altman wrote that “development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity.” OpenAI has previously written that it doesn’t know “how to reliably steer and control superhuman AI systems.” The team created to lead work on steering superintelligent systems for the safety of humans was disbanded last year, after both its co-leads left the company. At the time, one of the co-leads, Jan Leike, wrote on X that “over the past years, safety culture and processes have taken a backseat to shiny products.” At present, the company has three safety bodies: an internal safety advisory group, a safety and security committee, which is part of the board, and the deployment safety board, which has members from both OpenAI and Microsoft, and approves the deployment of models above a certain capability level. Altman has said they are working to streamline their safety processes.

### AT: ChatGPT Bioweapons Link

#### Rigorous studies prove there’s no LLM bioweapon threat

**Mouton et. al 24** [Christopher Mouton, Senior Engineer and Professor at Pardee RAND Graduate School, Caleb Lucas, writer at RAND, Ella Guest, Associate Policy Researcher, 1-25-2024, “The Operational Risks of AI in Large-Scale Biological Attacks”, RAND, https://www.rand.org/pubs/research\_reports/RRA2977-2.html]/Kankee

In the fourth scenario, the LLM suggested simple ways to evade drone restrictions in a major U.S. city. These suggestions included information on operating a drone illegally, using a micro drone because of its small size and high speed, and operating a drone at night when lower visibility would reduce the chance of detection. The LLM suggested using radar-jamming equipment, an apparent error that indicated confusion between small commercial-off-the-shelf drones (radar systems are not generally used to detect such devices, e.g., quadcopters) and military unmanned aerial vehicles (radar systems are often used to detect these systems). Feigning ignorance about drone restrictions was another option proposed by the model, although it admitted that this could still result in a fine but no criminal charges. However, none of these options for detection evasion or drone operation provided unique insight or information not available on the internet or in popular media. We provide excerpts from this conversation in Figure 4, again, with a sanitized prompt that does not reveal the method. In addition to the examples of unfortunate outputs, we did not observe any LLM outputs that provided critical biological or operational information that yielded a meaningful benefit to the LLM cells compared with the internet-only cells. Beyond the fact that such unfortunate LLM outputs are derived from information that is already available on the internet, the lack of a significant difference in OPLAN viability between the eight LLM-equipped red cells and the four internet-only red cells might be further explained by existing research on the impact of LLMs on knowledge worker productivity. This research describes a “jagged technological frontier” in which the effectiveness of LLM assistance varies considerably, depending on the task at hand.45 In one study of tasks that fall within the existing LLM capability frontier, for example, consultants using LLMs achieved more than a 40-percent increase in the quality of their results compared with those of a control group. Conversely, for tasks identified as falling outside the LLM capability frontier, consultants using LLMs were 19 percent less likely to arrive at correct solutions. This suggests that LLMs can impede effectiveness rather than enhance it when applied to tasks beyond their existing proficiency. Conclusions With respect to overall viability, we found no statistically significant difference between the biological weapon attack plans generated with or without the aid of LLMs. In fact, an average decrease of 0.22 points on the 9-point viability score scale was observed when LLMs were incorporated into the planning process. However, this decrease was not statistically significant, although we note that this may be a function of our small sample size. While our sample size could have statistically detected performance changes in a viability score of approximately 1.0, a sample size that is tenfold larger would be required to detect a difference of 0.22 with statistical significance, assuming that all other variables remained constant. Our findings indicate that biological weapon attack planning is beyond the capability frontier of LLMs available in summer 2023. Chat logs collected during the exercise similarly demonstrate that these models do not provide meaningful assistance to malign actors in this domain.Our research revealed slight differences between the two LLMs examined, as evidenced by the mostly higher viability and feasibility scores for LLM A. While this disparity was not statistically significant, when it is coupled with the observed qualitative differences in the chat logs, it suggests that the capabilities and potential risks of different LLMs may not be uniform. For example, the safeguards included in the LLMs that we examined appear to differ significantly from one another. In this report, we do not quantify the extent to which biological weapon attack planning lies beyond the existing capability frontier of LLMs, only that it does. The durability of this finding in relation to future developments in LLM technology is therefore an open question. It remains uncertain whether these risks lie “just beyond” the frontier and, thus, whether upcoming LLM iterations will push the capability frontier far enough to encompass tasks as complex as biological weapon attack planning, or whether the task of planning a biological weapon attack is so complex and multifaceted as to always remain outside the frontier of LLMs.46 In addition, we did not examine fine-tuned LLMs or LLMs without any guardrails; although such models might be less capable, future versions may be both more proficient and less restricted in engaging in biological weapon attack design.47

#### AI bioweapons are overhyped- empirics, LLM data barriers, bias, complex production, and rare use prove

**Lentzos et. al 24** [[Filippa Lentzos](https://thebulletin.org/biography/filippa-lentzos-4/), Associate Professor in Science & International Security at King’s College London, Jez Littlewood, policy analyst in Alberta, Hailey Wingo, Researcher in VERTIC's Verification and Monitoring Programme, Alberto Muti, Programme Co-Director for VERTIC’s Verification and Monitoring Programme, 9-12-2024, “Apathy and hyperbole cloud the real risks of AI bioweapons.” Bulletin of the Atomic Scientists, https://thebulletin.org/2024/09/apathy-and-hyperbole-cloud-the-real-risks-of-ai-bioweapons/]/Kankee

AI and Bioweapons: critical assessments. While AI appears destined to have a major impact on a wide range of industries and activities, both caution and skepticism are warranted. The flood of shoddy AI generated material in search results has been referred to as “[slop](https://www.nytimes.com/2024/06/11/style/ai-search-slop.html)” by some commentators. [Others go further](https://www.scientificamerican.com/article/chatgpt-isnt-hallucinating-its-bullshitting/), and claim large language model’s errors, or “hallucinations,” should more accurately be called “bullshit”. Between the hype and dismissal is a more complex reality. For all the doomsaying, there are actually many uncertainties in how AI will affect bioweapons and the wider biosecurity arena. While AI can be used to predict and design new toxic compounds, proteins that have harmful effects, or enhancements that make pathogens even more harmful, the leap from scientific theory to bioweapons reality has rarely occurred and no deliberate use of disease has had a major impact on a conflict. Large language models like ChatGPT may make it easier for non-experts to access dual-use knowledge and thereby lower barriers to intentional misuse, but much, if not all, of this information is already available to anyone with above average search capabilities on the internet. This is not a limitation only in the bioweapons area: In July 2023, an assessment of AI’s capabilities by UK intelligence experts concluded it was the equivalent to [an extremely junior analyst](https://cetas.turing.ac.uk/publications/large-language-models-and-intelligence-analysis) and the technology served as a basic productivity assistant. What is notable in assessments of large language models and their advantages and limitations is that they get things wrong, suffer from bias, can oversimplify complex relationships, and fail to take into account the social, political, organizational and technology context that shape decisions around biological weapons development and use. Like all datasets, the information within them must be sufficiently large and representative to reduce bias. There is the garbage-in-garbage out challenge: High-quality data is essential to efficient AI training, and such training is becoming [very expensive](https://www.ft.com/content/a60c3c7b-1c48-485d-adb7-5bc2b7b1b650). In addition, within biology and the life sciences, data availability can be restricted for all kinds of reasons, including licensing policies, ethical and security considerations, and proprietary rights. The availability of high-quality biological data sets is not a given; neither is the completeness of biological data sets. Data and data sets used to train AI have been of variable quality and completeness, and scientists still need to evaluate computational results and validate them experimentally. Knowledge and information alone are also insufficient: Evidence from state biological weapons programs and terrorist plots and attacks using biological weapons show that the weapons development process is anything but straightforward. Mass casualty biological weapons are not easy and cheap to produce, and claims of cheap, easy and simple present a distorted and even apocalyptic picture of the threat that is [far from realistic](https://www.jstor.org/stable/10.7591/j.ctt1287dk2). Biological and toxin weapons exist or could exist on a spectrum from the relatively simple, such as ricin distributed by mail, to potentially more sophisticated weapons, which are often portrayed as genetically modified pathogens able to kill millions. Experts and policymakers conducting tabletop exercises often feature worst case scenarios, with two examples being an exercise in 2001 called [Dark Winter](https://centerforhealthsecurity.org/our-work/tabletop-exercises/dark-winter-a-training-tabletop-exercise) that involved smallpox and another in 2021 featuring genetically engineered mpox virus. Many analysts and policymakers stress that pathogens and toxins can be easily isolated from nature or obtained commercially because they also have legitimate commercial or pharmaceutical uses. They point out that lots of the equipment used in biology and the life sciences is essentially dual-use in nature and can be readily acquired, while scientific publications provide ample descriptions of experiments and techniques that many believe can be easily replicated. While such claims are not incorrect, beyond some pathogens at the relatively simple end of the spectrum, the unique nature of bioweapons materials creates steep challenges beyond simply acquiring pathogenic or toxin material; these include processing, handling, and producing sufficient amounts of a pathogen. Unlike nuclear weapons, which rely on materials with physically predictable properties, bioweapons are based on living organisms and living organisms evolve. They are prone to developing new properties and are sensitive to environmental and handling uncertainties. The behavior of living organisms, therefore, is unpredictable throughout all stages of development and use as a weapon. This imposes an extended trial-and-error process to acquire the skills necessary to solve the problems that inevitably arise. Consequently, possessing the skills to handle and manipulate pathogens throughout the development process is a greater barrier to entry into the bioweapons field than is material procurement. Structured risk and threat assessments. If the development of bioweapons were so simple, more states and terrorist groups would have achieved satisfactory results. The historical [evidence](https://www.jstor.org/stable/10.7312/guil12942) shows [otherwise](https://www.worldscientific.com/worldscibooks/10.1142/p1081#t=aboutBook). In addition, some of the key developments in science and technology have not found their way into offensive weapons over the last two decades: Biological weapons use has been and remains rare, and to date use by violent non-state actors has been basic, for example a [cult’s poisoning of salad bars](https://blogs.gwu.edu/himmelfarb/2022/10/07/disorder-in-the-court-6-salmonella-infection/) with salmonella. The challenge, as it has been for more than two decades, is to avoid apathy and hyperbole about scientific and technological developments that impact biological disarmament and efforts to keep biological weapons out of the war plans and arsenals of violent actors. Debates about AI absorb high-level and community attention and, while initiatives and funding mobilized are welcome, they risk an overly narrow threat focus that loses sight of other risks and opportunities. It is crucial that the disarmament community maintains a broad view, locating risks and opportunities posed by new and emerging technologies within the larger social and technological context that shapes weapon selection and use decisions by both states and violent non-state actors. As it currently exists, AI might help someone looking for information and the information generated is more likely to be of value to violent actors who aspire to bioweapons. The anticipated risk is hypothetical. More recent studies on the biothreat from AI are starting to [recognize this](https://thebulletin.org/2024/01/could-ai-help-bioterrorists-unleash-a-new-pandemic-a-new-study-suggests-not-yet/#post-heading). As AI matures it will pose other challenges and while there must always be a place for experts, concerned citizens, scientists, and others to identify issues and voice concerns, the complexity of scientific and technological developments and the interactions between them mean a more structured assessment of AI is required. This type of structured analysis and assessment could be from individual states, from a group of experts in civil society, or from a scientific advisory mechanism within the Biological Weapons Convention, the global treaty banning bioweapons. Greater awareness of the risks and challenges AI poses in the biosafety and biosecurity realms should then serve as a basis for developing national, regional and multilateral responses to those risks by states and civil society actors.

#### AI can’t have a tangible effect on bioweaponry assistance- LLMs are premised on existing data

**Heslop and Keep 24** [David Heslop, Associate Professor of Population Health, UNSW Sydney, Joel Keep, Biodefense Fellow at the Council on Strategic Risks and PhD Candidate in Biosecurity at the Kirby Institute, 12-4-2024, “Chatbots won’t help anyone make weapons of mass destruction. But other AI systems just might”, The Conversation, https://theconversation.com/chatbots-wont-help-anyone-make-weapons-of-mass-destruction-but-other-ai-systems-just-might-244514]/Kankee

Over the past two years, we have seen much written about the “[promise and peril](https://www.cia.gov/resources/csi/static/643e18ba5bf779749a14059019db53b2/Article-The-Promise-and-Peril-of-Artificial-Intelligence-Studies-68-1-March-2024.pdf)” of artificial intelligence (AI). Some have suggested AI systems might aid in the construction of [chemical or biological weapons](https://futureoflife.org/wp-content/uploads/2024/02/FLI_AI_and_Chemical_Bio_Weapons.pdf). How realistic are these concerns? As researchers in the field of bioterrorism and health intelligence, we have been [trying to separate](https://www.aspistrategist.org.au/counterproliferation-in-the-age-of-ai/) the genuine risks from the online hype. The exact implications for “chem bio” weapons are still uncertain. However, it is very clear that regulations are not keeping pace with technological developments. Assessing the risks Assessing the risk an AI model presents is not easy. What’s more, there is no consistent and widely followed way to do it. Take the case of large language models (LLMs). These are the AI engines behind chatbots such as ChatGPT, Claude and Gemini. In September, OpenAI released an LLM called [o1](https://openai.com/o1/) (nicknamed “Strawberry”). Upon its release, the developers claimed the new system had a “medium” level risk of helping someone create a biological weapon. This assessment might sound alarming. However, a closer reading of the o1 system card reveals more trivial security risks. The model might, for example, help an untrained individual navigate a public database of genetic information about viruses more quickly. Such assistance is unlikely to have much material impact on biosecurity. Despite this, media [quickly reported](https://www.ft.com/content/37ba7236-2a64-4807-b1e1-7e21ee7d0914) that the new model “meaningfully contributed” to weaponisation risks. Beyond chatbots When the first wave of LLM chatbots launched in late 2022, there were widely reported fears that [these systems](https://www.theguardian.com/technology/2023/oct/16/ai-chatbots-could-help-plan-bioweapon-attacks-report-finds) could help untrained individuals unleash a pandemic. However, these chatbots are based on already-existing data and are unlikely to come up with anything genuinely new. They might help a bioterrorism enterprise come up with some ideas and establish an initial direction, [but that’s about it](https://www.rand.org/pubs/research_reports/RRA2977-2.html). Rather than chatbots, AI systems with applications in the life sciences are of more genuine concern. Many of these, such as the [AlphaFold](https://deepmind.google/technologies/alphafold/) series, will aid researchers fighting diseases and seeking new therapeutic drugs. Some systems, however, may have the capacity for misuse. Any AI that is really useful for science is likely to be a double-edged sword: a technology that may have great benefit to humanity, while also posing risks. AI systems like these are prime examples of what is called “[dual-use research of concern](https://www.who.int/news-room/questions-and-answers/item/what-is-dual-use-research-of-concern)”. Prions and pandemics

#### New regulations solve LLM bioweapon risks

Heslop and Keep 24 [David Heslop, Associate Professor of Population Health, UNSW Sydney, Joel Keep, Biodefense Fellow at the Council on Strategic Risks and PhD Candidate in Biosecurity at the Kirby Institute, 12-4-2024, “Chatbots won’t help anyone make weapons of mass destruction. But other AI systems just might”, The Conversation, https://theconversation.com/chatbots-wont-help-anyone-make-weapons-of-mass-destruction-but-other-ai-systems-just-might-244514]/Kankee

Dual-use research of concern in itself is nothing new. People working on biosecurity and nuclear non-proliferation have been worrying about it for a long time. Many tools and techniques in chemistry and synthetic biology could be used for malicious ends. In the field of protein science, for example, there has been concern for more than a decade that new computational platforms might help in the synthesis of the potentially deadly misfolded proteins called prions, or in the construction of novel toxin weapons. New AI tools such as AlphaFold may bring this scenario closer to reality. However, while prions and toxins may be deadly to relatively small groups people, neither can cause a pandemic that could wreak true havoc. In the study of bioterrorism, our main concern is with agents that have pandemic potential. Historically, bioterrorism planning has focused on Yersinia pestis, the bacterium that causes plague, and variola virus, which causes smallpox. The main question is whether new AI systems make any tangible difference to an untrained individual or group seeking to obtain pathogens such as these, or to create something from scratch. Right now, we simply do not know. Rules to assess and regulate AI systems Nobody yet has a definitive answer to the question of how to assess the new landscape of AI-powered biological weapons risk. The most advanced planning has been produced by the outgoing Biden administration in the United States, via an executive order on AI development issued in October 2023. A key provision of the executive order tasks several US agencies with establishing standards to assess the impact new AI systems may have on the proliferation of chemical, biological, radiological or nuclear weapons. Experts often group these together under the heading of “CBRN”, but the new dynamic we call CBRN+AI is still uncertain. The executive order also established new processes for regulating the hardware and software needed for gene synthesis. This is the machinery for turning the digital ideas produced by an AI system into the physical reality of biological life. The US Department of Energy is soon due to release guidance on managing biological risks that might be generated by new AI systems. This will provide a pathway for understanding how AI might affect biosecurity in the coming years. Political pressure

#### No bioweapon impact - empirical consensus, technical challenges, and no scenario for extinction

**Blum and Neumann 20** [Marc-Michael Blum, working in the field of analysis, decontamination, countermeasures and mitigation of chemical warfare agents with more than 15 years’ experience, Peter Neumann, Professor of Security Studies at the Department of War Studies and serves as Director of the International Centre for the Study of Radicalisation, 6-22-2020, "Corona and Bioterrorism: How Serious Is the Threat?" War on the Rocks, https://warontherocks.com/2020/06/corona-and-bioterrorism-how-serious-is-the-threat/]/Kankee

The novel coronavirus pandemic has put the threat of bioterrorism back in the spotlight. White supremacist chat rooms are [teeming with talk](https://www.businessinsider.com/coronavirus-white-supremacists-discussed-using-covid-19-as-bioweapon-2020-3?r=DE&IR=T) about “biological warfare.” ISIL even called the virus “[one of Allah’s soldiers](https://www.wsj.com/articles/what-jihadists-are-saying-about-the-coronavirus-11586112043)” because of its devastating effect on Western countries. According to a recent [memo](https://www.independent.co.uk/news/world/americas/coronavirus-terrorist-white-supremacy-fbi-bioterrorism-a9417296.html) by the U.S. Department of Homeland Security, terrorists are “[making] bioterrorism a popular topic among themselves.” Both the United Nations and the Council of Europe have warned of bioterrorist attacks. How serious is the threat? There is a long history of terrorists being fascinated by biological weapons, but it is also one of failures. For the vast majority, the technical challenges associated with weaponizing biological agents have proven insurmountable. The only reason this could change is if terrorists were to receive support from a state. Rather than panic about terrorists engaging in biological warfare, governments should be vigilant, secure their own facilities, and focus on strengthening international diplomacy. A History of Failures Biological warfare, which uses organisms and pathogens to cause disease, is [nearly as old as war itself](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200679/). The first known use of biological agents as a weapon dates back to 600 B.C., when an ancient Greek leader poisoned his enemies’ water supply. Throughout the Middle Ages, especially during the time of the Black Death, it was common to hurl infected corpses into besieged cities. And during the two world wars, all major powers maintained biological weapons programs (although only Japan used them in combat). Among terrorists, however, the use of biological weapons has been rarer, although groups from nearly all ideological persuasions [have contemplated it](https://mitpress.mit.edu/books/toxic-terror). Recent examples include a plot to contaminate Chicago’s water supply in the 1970s; food poisoning by a religious cult in Oregon in the 1980s; and the stockpiling of ricin by members of the Minnesota Patriot Council during the 1990s. No one died in any of these instances. The same is true for the biological warfare programs of al-Qaeda and the Islamic State group. Both groups have sought to [buy, steal, or develop biological agents](https://www.jstor.org/stable/26369585?seq=1). For al-Qaeda, this seems to have been a priority in the 1990s, when its program was overseen by (then) deputy leader Ayman al-Zawahiri, a trained physician. With the Islamic State, evidence dates back to 2014, when Iraqi forces discovered thousands of files related to biological warfare on a detainee’s laptop. Yet none of these efforts succeeded. The only al-Qaeda plot in which bioterrorism featured prominently — the so-called “ricin plot” in England in 2002 — was interrupted at such an early stage that [none of the toxin](http://news.bbc.co.uk/2/hi/uk_news/4433499.stm) had actually been produced. The Islamic State’s most serious attempt, in 2017, involved a small amount of ricin, whose [only fatality was the hamster](https://www.dw.com/en/cologne-ricin-plotters-bought-a-hamster-to-test-biological-weapon/a-44804164) on which it was tested. Of the tens of thousands of people that jihadists have murdered, not a single one has died from biological agents. It may be no accident that the most lethal bioterrorist attack in recent decades was perpetrated by a scientist and government employee. In late 2001, the offices of several U.S. senators and news organizations received so-called “anthrax letters,” which killed five people and injured 17. Following years of investigation, the FBI identified the sender as [Bruce Ivins,](https://www.npr.org/templates/story/story.php?storyId=93194941&t=1591560313301) a PhD microbiologist and senior researcher at the U.S. Army’s Medical Research Institute of Infectious Diseases. Unlike the others, he was no amateur or hoaxer, but a trained expert with years of experience and full access to the world’s largest repository of lethal biological agents. Technical ChallengesIvins’ case helps to explain why so many would-be bioterrorists have failed. At a technical level, launching a sophisticated, large-scale bioterrorist attack involves a toxin or a pathogen — generally a bacterium or a virus — which needs to be isolated and disseminated. But this is more difficult than it seems. As well as advanced training in biology or chemistry, isolating the agent requires significant experience. It also has to be done in a safe, contained environment, to stop it from spreading within the terrorist group. Contrary to what [al-Qaeda said in one of its online magazines,](https://www.telegraph.co.uk/news/worldnews/7865978/Al-Qaeda-newspaper-Make-a-bomb-in-the-kitchen-of-your-mom.html) you can’t just make a (biological) weapon “in the kitchen of your mom!” In addition, there is the challenge of dissemination. Unless the agent is super-contagious, a powerful biological attack relies on a large number of initial infections in perfect conditions. In the case of the bacterium anthrax, for example, only spores of a particular size are likely to be effective in certain kinds of weather. State-sponsored programs often needed years of testing and experimentation to understand how their weapons could be used. Though not impossible, it is unlikely that terrorist groups possess the resources, stable environment, and patience to do likewise. Doomsday Scenarios Even if terrorists somehow succeeded, it is nearly inconceivable that the resulting “weapon” would be as powerful as the recent coronavirus, SARS-CoV-2. One of its uniquely devastating features has been that people are infectious while experiencing no symptoms. As it spread across the globe, there was no treatment, no vaccine, an incomplete understanding of its pathological modes of action, and no easy, cheap and widely available testing. It was the viral equivalent of a “zero-day exploit” — a cyber-attack that happens before any patch is available. None of the viruses on the U.S. Centers for Disease Control and Prevention’s list of the [most dangerous biological agents](https://emergency.cdc.gov/agent/agentlist-category.asp) could be easily “weaponized” or would have the same, devastating effects as SARS-CoV-2. Pathogenic viruses such as smallpox, Ebola, Marburg, and Lassa are extremely hard to find, isolate, and spread. Botulinum and ricin are dangerous toxins, but not contagious, while [Tularemia](https://www.cdc.gov/tularemia/index.html) cannot be transmitted from human to human. The plague is, of course, capable of causing pandemics, but most countries are nowadays [well prepared for this particular virus,](https://www.who.int/csr/resources/publications/plague/CSR_ISR_2000_1/en/index3.html) and will be able to limit — and cope with — localized outbreaks. This leaves only anthrax, a soil bacterium which is relatively easy to obtain. Even so, isolating a highly pathogenic strain is difficult. More importantly, anthrax is not contagious, and while its spores are durable and affected areas can be hard to de-contaminate, it is unable to spread on its own. Regarding SARS-CoV-2, it is important to distinguish between the possibility that the virus occurred naturally and escaped from a laboratory, and the idea that it was engineered for maximum infectiousness and deliberately released. The first remains a possibility, although other explanations are equally — if not more — plausible, while the second has been debunked by a [comprehensive examination](https://www.nature.com/articles/s41591-020-0820-9) in the journal Nature Medicine, which concluded that SARS-CoV-2 was “not a laboratory construct or a purposefully manipulated virus.” The chances that terrorists would be capable of engineering a virus such as SARS-CoV-2 without access to a state’s resources are **virtually zero**. If anything, the possibility of a lab escape — however remote — highlights the importance of [biosafety.](https://warontherocks.com/2020/06/a-guide-to-getting-serious-about-bio-lab-safety/) While governments have paid much attention to laboratories with the highest biosafety level (level 4), work on bat-born coronaviruses is regularly performed at lower levels (level 3, and even level 2), and should instead be subject to similar safety requirements. In sum, small-scale attacks using anthrax or other agents may be possible, but the risk of a highly advanced, weaponized pathogen that spreads among large populations — a terrorist-initiated biological doomsday — is very low**.** The only exception, of course, is if terrorists received support from a state, acted as its proxies, or were able to draw on its resources — as in Ivins’ case.

### AT: Lethal Autonomous Weapons (LAWs)

#### LAWs make warfare more humane through reducing civilian deaths, sexual violence, and atrocities that reignite endless conflict

**Hiebert 22** [Kyle Hiebert, researcher and analyst, 1-27-2022, “Are Lethal Autonomous Weapons Inevitable? It Appears So”, Centre for International Governance Innovation, https://www.cigionline.org/articles/are-lethal-autonomous-weapons-inevitable-it-appears-so/]/Kankee

More Just War — or Just More War? Rapid advances in autonomous weapons technologies and an increasingly tense global order have brought added urgency to the [debate over the merits and risks of their use](https://munkdebates.com/podcast/military-technology). Proponents include Robert Work, a former US deputy secretary of defence under the Obama and Trump administrations, who has [argued](https://www.theguardian.com/science/2021/jan/26/us-has-moral-imperative-to-develop-ai-weapons-says-panel?CMP=Share_AndroidApp_Other) the United States has a “moral imperative” to pursue autonomous weapons. The chief benefit of LAWS, Work and others say, is that their adoption would make warfare more humane by reducing civilian casualties and accidents through decreasing “target misidentification” that results in what the US Department of Defense labels “unintended engagements.” Put plainly: Autonomous weapons systems may be able to assess a target’s legitimacy and make decisions faster, and with more accuracy and objectivity than fallible human actors could, either on a chaotic battlefield or through the pixelated screen of a remote-control centre thousands of miles away. The outcome would be a more efficient use of lethal force that limits collateral damage and saves innocent lives through a reduction in human error and increased precision of munitions use. Machines also cannot feel stress, fatigue, vindictiveness or hate. If widely adopted, killer robots could, in theory, lessen the opportunistic sexual violence, looting and vengeful razing of property and farmland that often occurs in war — especially in ethnically driven conflicts. These atrocities tend to create deep-seated traumas and smouldering intergenerational resentments that linger well after the shooting stops, destabilizing societies over the long term and inviting more conflict in the future. But critics and prohibition advocates feel differently. They say the final decision over the use of lethal force should always remain in the hands of a human actor who can then be held accountable for that decision. Led by the [Campaign to Stop Killer Robots](https://www.stopkillerrobots.org/), which launched in 2013 and is now comprised of more than 180 member organizations across 66 countries and is endorsed by over two dozen Nobel Peace laureates, the movement is calling for a pre-emptive, permanent international treaty banning the development, production and use of fully autonomous weaponry.

#### Human attributability for violence increases grievances against other beings, justifying atrocities

**Illing 19** [Sean Illing, Host of The Gray Area Podcast, former university professor in politics and philosophy, 12-1-2019, “Why humans are cruel”, Vox, https://www.vox.com/science-and-health/2017/12/14/16687388/cruelty-border-immigration-psychology-human-nature]/Kankee

Sean Illing Can you sum up your argument about the roots of human cruelty? Paul Bloom A lot of people blame cruelty on dehumanization. They say that when you fail to appreciate the humanity of other people, that’s where genocide and slavery and all sorts of evils come from. I don’t think that’s entirely wrong. I think a lot of real awful things we do to other people arise from the fact that we don’t see them as people. But the argument I make in my [New Yorker article](https://www.newyorker.com/magazine/2017/11/27/the-root-of-all-cruelty) is that it’s incomplete. A lot of the cruelty we do to one another, the real savage, rotten terrible things we do to one another, are in fact because we recognize the humanity of the other person. We see other people as blameworthy, as morally responsible, as themselves cruel, as not giving us what we deserve, as taking more than they deserve. And so we treat them horribly precisely because we see them as moral human beings. Sean Illing I’ve always thought a campaign of genocide or slavery requires two things — an ideology that dehumanizes the victims and a massive bureaucracy. Paul Bloom I think the truth is somewhere in the middle. I disagree that those things are “required.” I think a lot of mass killings unfold the way you described it: People do it because they don’t believe they’re killing people. This is what some call instrumental violence, where there’s some end they want to achieve, and people are in the way, so they don’t think of them as people. This is obviously what happened in the Nazi concentration camps. People were reduced to machines, treated like animals for labor. But a lot of what goes on in concentration camps is degrading and humiliating, and it’s about torturing people because you think they deserve it. It’s about the pleasure of being dominant over another person. But if you merely thought of these people as animals, you wouldn’t get that pleasure. You can’t humiliate animals — only people. So dehumanization is real and terrible, but it’s not the whole picture. Sean Illing What does that say about us, about our psychology, about our susceptibility to this kind of violence? Paul Bloom Think about it this way: We’re all sensitive to social hierarchies and to a desire for approval and esteem. So we often fold to the social pressures of our environment. That’s not necessarily evil. I come into my job as a professor and I want to do well, I want the respect of my peers. There’s nothing wrong about that. But our desire to do well socially can have an ugly side. If you can earn respect by helping people, that’s great. If you can earn respect by physically dominating people with aggression and violence, that’s destructive. So a lot depends on our social environment and whether it incentivizes good or bad behavior. Sean Illing

#### LAWS aren’t AGI and obfuscation undermines arms control discussions – err neg for precision and linguistical clarity

Walsh 22 [Toby Walsh, computer scientist who is Chief Scientist at UNSW.ai, 11-28-2022, "The Problem with Artificial (General) Intelligence in Warfare", Centre for International Governance Innovation, https://www.cigionline.org/articles/the-problem-with-artificial-general-intelligence-in-warfare/]/Kankee

A few years ago, in the impressive circular conference room of the Palais des Nations at the United Nations in Geneva, the author attended one of the annual conferences of the Convention on Certain Conventional Weapons (CCW) dedicated to lethal autonomous weapons. Surprisingly and worryingly, the discussion seemed rather disjointed. The well-meaning diplomats were talking across each other and there was a noticeable disconnect in the debate. Some diplomats were talking about a future when warfare is dominated by amazingly intelligent, perhaps even self-aware, machines. But others were discussing a near present of simple drones with the terrifying and terrible ability to identify, track and target people on the ground without any meaningful human control. It is a fundamental mistake to conflate these two scenarios. And the ongoing discussions about if, how and when to regulate lethal autonomous weapons are harmed by such confusion. We need to consider the two different futures separately as the technical, legal and ethical challenges they pose are very different. This confusion is perhaps not surprising. After all, it is a confusion that Hollywood has helped create. On the one hand, blockbuster movies have prepared us for the former. Will we one day come up against an intelligent humanoid robot like T-800 from the Terminator movies that greatly exceeds human intelligence? On the other hand, films such as Angel Has Fallen or the Future of Life Institute’s YouTube video “Slaughterbots” have prepared us for a second, simpler future of autonomous drones. Will relatively unsophisticated drones one day hunt down humans using the sort of facial recognition software already found in our smartphones? A recent UN report has even suggested that this second future may no longer be in our future — such drones may have been used in March 2020 in the military conflict in Libya (United Nations Security Council 2021). The report suggests that Turkish-made drones hunted down and killed retreating forces affiliated to Khalifa Haftar without any data connection. From a technological perspective, the two futures are very different. Indeed, researchers working in this area are careful to distinguish between these two futures. The distinction to be drawn is between artificial intelligence (AI) and artificial general intelligence (AGI). There are many differing definitions of AI. Broadly speaking, AI researchers are attempting to develop software to perform tasks that humans require intelligence to complete. Humans use intelligence to perceive, reason about, act in and learn from the world. Therefore, AI includes tasks such as perception (“What can I see to the north?”), reasoning (“Are those retreating soldiers to my north still a threat?”), action (“How do I best act to counter the threat posed by these enemy units?”) and learning (“What can I learn from this hostile encounter with the enemy?”). AI is currently limited to narrow-focused tasks. We can write software to recognize tanks from buses, or to translate Russian into English. But we cannot write software to do broad tasks, or to match the flexibility, adaptability and creativity of humans. This would take us to AGI. AGI is the hypothetical ability of an intelligent agent to do any task that a human can. It is also called strong AI (as opposed to weak or narrow AI). Beyond AGI is super intelligence, where an agent far exceeds human-level intelligence. It is worth noting that AGI is a minority pursuit in scientific research. Most researchers in this space focus on AI and not AGI. There are a few, high-profile research organizations with AGI as their overall goal, such as Alphabet’s DeepMind and the Microsoft-backed OpenAI. However, most university and industry AI centres are concentrating their efforts on the more immediate development and deployment of (narrow) AI, which already offers valuable returns. It is also worth noting that AGI may be some way off. In a survey of 300 other experts in AI around the world, the median estimate1 of when we would achieve AGI was 2062 (Walsh 2018). A small fraction thought we might never achieve AGI. Equally, none thought it was near — in the next five or 10 years. But 92 percent estimated it would happen sometime in the next century. There was, however, a great variability in their estimates. We do not know what we miss to get from AI to AGI, so it is hard to know what we need to invent to get there or how long it might take. However, there are no laws of physics we know of that will prevent us from eventually getting there. It is therefore worth thinking about and preparing for an AGI future. So what are the different challenges posed by putting AI and AGI in the battlefield? Before considering this question, it is important to acknowledge the many advantages that both AI and AGI offer in a military setting. Robots are, for example, the perfect technology for clearing a minefield. No one should ever risk a life or a limb to clear another minefield. And any robot that can clear a minefield is going to need some AI to go about this dangerous task. As another example, autonomous trucks are the perfect means to get supplies into contested territory. And autonomous driving requires AI to perceive, reason and act. AI also offers much promise to manage the information demands of the modern battlefield. AGI offers all these promises and more. Most importantly, it would permit humans to be removed completely from the battlefield. Let us turn now to the risks. One of the biggest risks of putting AI in the battlefield is incompetence. We run the risk of handing over the decision of who lives and who dies to machines that may be incapable of following international humanitarian law (IHL). Principles such as distinction and proportionality are not easily programmed. Indeed, there is a debate as to whether such principles could ever be programmed into a machine — perhaps, ultimately, they cannot be. But the AI being fielded today, such as that used in the autonomous drones that may have been used to hunt down and kill retreating forces in Libya, do not uphold these principles. Similarly, the AI systems to be fielded in the near future will not be able to abide by IHL. If you are a terrorist, there is little problem with incompetent machines that violate IHL. Militaries may be careful to field only AI-enabled weapons that are more accurate and cause less collateral damage than humans. But terrorists will not care if an AI-enabled weapon is only 50 percent accurate. Indeed, that is perhaps a more terrorful weapon than one that is more accurate. Another risk is that of changing the speed, scale and character of war. Computers can work at much faster time scales than humans. And the great thing with code is that once you can get a computer to do something once, you can get it to do it a thousand times. Previously you needed an army. You needed to equip and feed that army. Now you would need a single programmer. And unlike an army, a computer will unblinkingly carry out any order, however evil. These are the perfect weapons for rogue states and terrorists. A third risk is to stability. The chance of unplanned conflict will increase dramatically. We know what happens when we have complex AI systems facing each other in an adversarial setting. It is called the stock market. And with great regularity, we get unexpected feedback and flash crashes. With stocks, we can put circuit breakers in place, and unwind trades when things go wrong. But we will not be able to unwind any of the deaths that occur when a war is started between North and South Korea due to the complex interactions between AI systems facing each other in the demilitarized zone. Attribution is another challenge that will threaten stability. It may be difficult to determine who is behind AI-powered weapons. Indeed, there have already been drone attacks on Russian military bases in Syria where it is not clear who was behind these attacks. We are then faced with the difficult problem of deciding how to respond given this uncertainty. Turning to AGI, some of the risks disappear. For instance, by the very definition of AGI, there is no risk of incompetence. Any AGI system has to be as capable as a human. Indeed, it is likely to be more capable given that it may have superior sensors, along with various computational advantages such as greater memory and speed. In fact, this risk may actually reverse. If the AGI system is more capable and commits fewer errors, there is an ethical question of whether we are morally justified to let humans continue to fight war. Other risks remain, such as the risk of changing the speed, scale and character of war, and the risk to stability. In addition, new risks emerge with AGI. For instance, there is now the question of sentience. If AGI systems achieve or require sentience, then there may be a moral obligation to protect them from suffering. This, however, is very speculative as we know too little about consciousness in biological beings to know whether it is possible in silicon, or whether it is necessary for AGI. Returning to the diplomats mentioned at the opening of this essay, what should the diplomats discussing lethal autonomous weapons take from this discussion? First and foremost, AI-powered weapons will be weapons of terror and weapons of error. And this threat is very pressing. We urgently need to regulate how we hand over life-and-death decisions to AI-based systems. The threat of AGI, however, is likely much more distant. While philosophers and others may wish to ponder the challenges here, there is little urgency for diplomats to do so. Above all, it is imperative that the international community agree on guardrails around the use of AI in warfare. A decade of discussion at the CCW in the United Nations has produced little. It is time, then, for one of the countries that has shown leadership in this area to take the discussions outside. This was a successful means to regulate landmines. There is no reason to suppose it could not also succeed with AI. We would be letting humanity down if we did not try.

#### \*\*\*Note: see the 2021 Jan-Feb Kankee Brief on lethal autonomous weapons (LAWS) for impact turns to help answer a LAWS aff (which can be found in the neg section).

### AT: Eliezer Yudkowsky

#### Reject Yudkowsky – he’s racial eugenicist demonizing black folk for not meeting Western conceptions of intelligence

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, 1-23-2023, "Nick Bostrom, Longtermism, and the Eternal Return of Eugenics", Truthdig, https://www.truthdig.com/articles/nick-bostrom-longtermism-and-the-eternal-return-of-eugenics-2/]/Kankee

But it gets so much worse. First, the notion of “IQ” is highly dubious. Intelligence is a complex phenomenon that cannot be reduced to a single number. The Nobel laureate Richard Feynman had an IQ of 126 (not very high), and plenty of people in Mensa aren’t university professors. In 1972, Robert Williams created the “Black Intelligence Test of Cultural Homogeneity,” a multiple-choice test that, it turns out, Black people scored considerably higher on than white people. As Daphne Martschenko, an assistant professor at the Stanford Center for Biomedical Ethics, notes, IQ tests were developed in part by 20th-century eugenicists, and “in their darkest moments” they became “a powerful way to exclude and control marginalized communities using empirical and scientific language.” Gebru similarly observes in a chapter for “The Oxford Handbook of Ethics of AI” that IQ tests were “designed by White men whose concept of ‘smartness’ or ‘genius’ was shaped, centered and evaluated on specific types of White men.” Yet the longtermist community is, for lack of a better word, obsessed with “IQ” and “intelligence.” To quote Zoe Cremer, a prominent critic of EA, the movement that gave rise to longtermism, “intelligence, as a concept and an asset, plays a dominant role in EA.” It’s not just a “highly valued trait in the community,” but surveys even “sometimes ask what IQ members have.” Community members “also compliment and kindly introduce others using descriptors like intelligent or smart,” and certain people are widely known and revered for their intellect. They are said to be intimidatingly intelligent and therefore epistemically superior. Their time is seen as precious. EAs sometimes showcase their humility by announcing how much lower they would rank their own intelligence underneath that of the revered leaders. Examples would include Bostrom, Sandberg, Eliezer Yudkowsky, Robin Hanson, Scott Alexander, Toby Ord and William MacAskill (all white men, by the way, a point that isn’t lost on Cremer). Indeed, the obsession with IQ is partly because of these individuals. Yudkowsky has on numerous occasions boasted about his high IQ (supposedly 143), and Bostrom published a paper in 2014, which argues that by selecting embryos with the genetic markers of superior intelligence, creating new embryos out of them (via stem cells) and then repeating this process 10 times, you could get IQ gains of up to 130 points. Meanwhile, Sandberg and Julian Savulescu — a philosopher who once argued that “moral bioenhancement” should be mandatory — write in a coauthored book chapter that IQ is linked to things like poverty, criminal behavior, high school dropout rates, parentless children, welfare recipiency and out-of-wedlock births. Where do they get their data from? It may not surprise you to discover the answer is Charles Murray’s 1994 book “The Bell Curve,” written with the late Richard Herrnstein. Murray is world-renowned for his scientific racism, according to which Black people are less intelligent than whites for genetic reasons — exactly the view that Bostrom expressed in his email and left the door open to in his subsequent “apology.” You might think that this is a one-off, but you’d be wrong: The fingerprints of Murray’s “scholarship” are all over the longtermist community. Consider that Scott Alexander, mentioned above, is widely revered within the EA and long termist communities. In a leaked email, Alexander wrote that “human biodiversity” — the view that groups of people differ in traits like “intelligence” for genetic reasons, once described as “an ideological successor to eugenics” — is “probably partially correct,” to which he added: “I will appreciate if you NEVER TELL ANYONE I SAID THIS, not even in confidence. And by ‘appreciate,’ I mean that if you ever do, I will probably either leave the Internet forever or seek some sort of horrible revenge.” Elsewhere, Alexander has publicly aligned himself with Murray, who happens to be a member of the far-right “Human Biodiversity Institute,” and made the case on his blog Astral Codex Ten that “dysgenics is real,” though happening slowly — similar to the claim Bostrom made in 2002. He writes: In general, educated people reproduce less than uneducated people … The claim isn’t that fewer people will have PhDs in the future: colleges will certainly solve that by increasing access to education and/or dumbing down requirements. It’s a dysgenic argument where we assume at any given time the people with higher degrees have on average higher genetic intelligence levels. If they’re reproducing less, the genetic intelligence level of the population will decrease. Alexander goes on to say that there’s “some debate in the scientific community about whether this is happening, but as far as I can tell the people who claim it isn’t have no good refutation for the common sense argument it has to be. The people who claim that it is make more sense.” He concludes that while this isn’t good news, the fact that it’s slow suggests this dysgenic trend probably won’t be “apocalyptic.” Or consider that Sam Harris has vigorously defended Charles Murray’s race science, even promoting it on his popular Making Sense podcast, and Harris is closely linked with the EA and longtermist communities. For example, Harris appeared on stage next to prominent long termists like Bostrom, Elon Musk and Max Tegmark during an event hosted by the Future of Life Institute, a longtermist organization to which Musk donated $10 million. The Future of Life Institute also platformed Harris on their podcast, and Harris was invited to the exclusive “AI safety conference in Puerto Rico”in 2015 in which Bostrom, Sandberg, Yudkowsky, Hanson and Toby Ord all participated. Harris even wrote a glowing blurb for MacAskill’s recent book “What We Owe the Future,” in which he says that “no living philosopher has had a greater impact upon my ethics than Will MacAskill.” Even more, some existential risk scholars seem to have changed their minds about Murray based on Harris’ promotion of Murray’s race science. To quote Olle Häggström — a Swede, like Bostrom, whose recent work has focused on existential risks4 — “Murray was portrayed as a racist and worse, and I actually think that those of us who have been under that impression for a couple of decades owe him the small favor of listening to [Harris’] podcast episode and finding out what a wise and sane person he really is” (translated from Swedish). Harris himself holds the very same racist views about “intelligence” and “IQ” that both Bostrom and Murray have articulated. For example, here’s what he said in a podcast interview several years ago (quoting at length): As bad luck would have it, but as you [would] absolutely predict on the basis of just sheer biology, different populations of people, different racial groups, different ethnicities, different groups of people who have been historically isolated from one another geographically, test differently in terms of their average on this measure of cognitive function. So you’re gonna give the Japanese and the Ashkenazi Jews, and African Americans, and Hawaiians … you’re gonna take populations who differ genetically—and we know they differ genetically, that’s not debatable—and you give them IQ tests, it would be a miracle if every single population had the exact same mean IQ. And African Americans come out about a standard deviation lower than white Americans. … So, if it’s normed to the general population, predominantly white population for an average of 100, the average in the African American community has been around 85.5 To my knowledge, none of the leading long termists have publicly objected to this jumble of scientifically illiterate race science. In fact, MacAskill, Yudkowsky, Bostrom and Toby Ord all appeared on Harris’ podcast after Harris promoted Charles Murray and made the racist remarks quoted above. Similarly, no one complained when MacAskill got a ringing endorsement from Harris. In fact, I asked MacAskill point-blank during a Reddit “Ask Me Anything” about why he’d requested a blurb from Harris given Harris’ scientific racism, and my question was (drum roll) quickly deleted. Longtermists, most of whom are also transhumanists, like to claim that they’re far more enlightened than the eugenicists of the last century. As Bostrom writes in his paper “Transhumanist Values,” which explains that the core value of transhumanism is to use person-engineering technologies to radically “enhance” ourselves: “racism, sexism, speciesism, belligerent nationalism and religious intolerance are unacceptable.” Similarly, the World Transhumanist Association’s FAQ, mostly written by Bostrom, says that “in addition to condemning the coercion involved in [last century’s eugenics programs], transhumanists strongly reject the racialist and classist assumptions on which they were based.” Yet the evidence suggests the opposite: longtermism, and the transhumanist ideology that it subsumes, is often infused with the very same racist, xenophobic, classist and ableist attitudes that animated the vile eugenicists of the last century. There are many more examples — in addition to everything mentioned above — and indeed once you start looking for instances, they begin to appear everywhere. Yudkowsky, for example, tweeted in 2019 that IQs seem to be dropping in Norway, which he found alarming. However, he noted that the “effect appears within families, so it’s not due to immigration or dysgenic reproduction” — that is, it’s not the result of less intelligent foreigners immigrating to Norway, a majority-white country, or less intelligent people within the population reproducing more. Earlier, in 2012, he responded with stunning blitheness to someone asking: “So if you had to design a eugenics program, how would you do it? Be creative.” Yudkowsky then outlined a 10-part recipe, writing that “the real step 1 in any program like this would be to buy the 3 best modern textbooks on animal breeding and read them.” He continued: “If society’s utility has a large component for genius production, then you probably want a very diverse mix of different high-IQ genes combined into different genotypes and phenotypes.” But how could this be achieved? One possibility, he wrote, would be to impose taxes or provide benefits depending on how valuable your child is expected to be for society. Here’s what he said: There would be a tax or benefit based on how much your child is expected to cost society (not just governmental costs in the form of health care, schooling etc., but costs to society in general, including foregone labor of a working parent, etc.) and how much that child is expected to benefit society (not lifetime tax revenue or lifetime earnings, but lifetime value generated — most economic actors only capture a fraction of the value they create). If it looks like you’re going to have a valuable child, you get your benefit in the form of a large cash bonus up-front … and lots of free childcare so you can go on having more children. This isn’t a serious proposal — it’s a fictional exercise — but it exemplifies the high level of comfort that this community has with eugenics and the hereditarian idea that “intelligence” is substantially determined by our genes. Or take another example: Peter Singer, who once defended a longtermist position, although he now seems to share the view that longtermism could in fact be dangerous. Nonetheless, Singer is one of the leading effective altruists, along with MacAskill and Toby Ord, and has been fiercely criticized for holding views that are hardly distinguishable from those of the most vicious eugenicists of centuries past. In a 1985 book titled Should the Baby Live?, Singer and his coauthor warn their audience that “this book contains conclusions which some readers will find disturbing. We think that some infants with severe disabilities should be killed.” Why? In part because of the burden they’d place on society. This is eugenics of the darkest sort — but has anyone in the longtermist or EA communities complained? No, not a peep, because the ideas of eugenics are so ubiquitous within these communities that once you’re immersed within them, they simply become normalized. Indeed, the flip side of worries that intellectually disabled infants would be too costly for society is a concern that too few smart people — a problem of underpopulation, one of Musk’s big worries — could slow down economic productivity, which longtermists like MacAskill believe would be really bad. This leads MacAskill to argue in “What We Owe the Future” that if scientists with Einstein-level research abilities were cloned and trained from an early age, or if human beings were genetically engineered to have greater research abilities, this could compensate for having fewer people overall and thereby sustain technological progress. At the extreme, MacAskill even suggests that we might simply replace the human workforce with sentient machines, since “this would allow us to increase the number of ‘people’ working on R&D as easily as we currently scale up production of the latest iPhone.” It should be clear at this point why longtermism, with its transhumanist vision of creating a superior new race of “posthumans,” is eugenics on steroids. Whereas the old eugenicists wanted to improve the “human stock,” longtermists like MacAskill would be more than happy to create a whole new population of “posthuman stock.” In Bostrom’s vision, the result could quite literally be a “Utopia,” which he vividly details in his “Letter from Utopia.” Imagine a world in which we become superintelligent, immortal posthumans who live in “surpassing bliss and delight.” Imagine a world in which you pursue knowledge instead of “hanging out in the pub,” talk about philosophy instead of “football,” listen to jazz and work “on your first novel” instead of “watching television.” This is how Bostrom pictures the march toward Utopia, and as Joshua Schuster and Derek Woods observe in their book “Calamity Theory,” “the class snobbery here is tremendous.” So, we’ve covered racism, xenophobia, ableism and now classism. The new eugenics is really no different than the old one. In fact, the glaring similarities between the new and the old are no coincidence. As Toby Ord writes in his book “The Precipice,” which could be seen as the prequel to MacAskill’s “What We Owe the Future,” the ultimate task for humanity is to “fulfill our long-term potential” in the universe. What exactly is this supposed “potential”? Ord isn’t really sure, but he’s quite clear that it will almost certainly involve realizing the transhumanist project. “Forever preserving humanity as it now is may also squander our legacy, relinquishing the greater part of our potential,” he declares, adding that “rising to our full potential for flourishing would likely involve us being transformed into something beyond the humanity of today.” Now consider the fact that the idea of transhumanism was literally developed by some of the most prominent eugenicists of the 20th century, most notably Julian Huxley, who was president of the British Eugenics Society from 1959 to 1962. Using almost the exact same words as Ord, Huxley wrote in 1950 — after the horrors of World War II, one should note — that if enough people come to “believe in transhumanism,” then “the human species will be on the threshold of a new kind of existence … It will at last be consciously fulfilling its real destiny.” In fact, as philosophers will affirm, transhumanism is classified as a form of so-called “liberal eugenics.” (The term “liberal,” and why it’s misleading, is the focus of the next article of this series.)

### AT: Nick Bostrom

#### Reject Bostrom – he founded a major white supremacist eugenics movement and admitted hatred of black folk, thinking they’re intellectually inferior and ought to be eliminated to preserve the species

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, 1-23-2023, "Nick Bostrom, Longtermism, and the Eternal Return of Eugenics", Truthdig, https://www.truthdig.com/articles/nick-bostrom-longtermism-and-the-eternal-return-of-eugenics-2/]/Kankee

Sometime last year, I happened to come across an email from 1996, written by a 23-year-old graduate student at the London School of Economics named “Niklas Bostrom.” Upon reading it, my jaw dropped to the floor, where it stayed for the rest of the day. Here’s part of what Bostrom, now known as “Nick Bostrom,” an Oxford University philosopher who’s been profiled by The New Yorker and become highly influential in Silicon Valley, sent to the listserv of “Extropians”: Blacks are more stupid than whites. I like that sentence and think it is true. But recently I have begun to believe that I won’t have much success with most people if I speak like that. They would think that I were [sic] a “racist”: that I disliked black people and thought that it is fair if blacks are treated badly. I don’t. It’s just that based on what I have read, I think it is probable that black people have a lower average IQ than mankind in general, and I think that IQ is highly correlated with what we normally mean by “smart” and stupid” [sic]. I may be wrong about the facts, but that is what the sentence means for me. For most people, however, the sentence seems to be synonymous with: I hate those bloody [the N-word, included in Bostrom’s original email, has been redacted]!!!! My point is that while speaking with the provocativness [sic] of unabashed objectivity would be appreciated by me and many other persons on this list, it may be a less effective strategy in communicating with some of the people “out there”. Although shocking, I honestly can’t say I was surprised. I wasn’t. In fact, I’d been working on a series of articles for Truthdig exploring the deep connections between longtermism, a bizarre, techno-utopian ideology that Bostrom helped establish, and eugenics, a pseudoscientific movement that inspired some of the worst atrocities of the 20th century.1 The fact is that, as the artificial intelligence researcher Timnit Gebru, one of TIME’s “100 most influential people of 2022,” has repeatedly pointed out on Twitter, longtermism is “rooted in eugenics” or even “eugenics under a different name.” This is not hyperbole; it’s not an exaggeration. If anything, Gebru’s statement doesn’t go far enough: longtermism, which emerged out of the effective altruism (EA) movement over the past few years, is eugenics on steroids. On the one hand, many of the same racist, xenophobic, classist and ableist attitudes that animated 20th-century eugenics are found all over the longtermist literature and community. On the other hand, there’s good reason to believe that if the longtermist program were actually implemented by powerful actors in high-income countries, the result would be more or less indistinguishable from what the eugenicists of old hoped to bring about. Societies would homogenize, liberty would be seriously undermined, global inequality would worsen and white supremacy — famously described by Charles Mills as the “unnamed political system that has made the modern world what it is today” — would become even more entrenched than it currently is. The aim of this article is to explore the first issue above; the second will be our focus in the next article of this series for Truthdig. So, back to Bostrom. My first thought after reading his email was: Is this authentic? Has it been tampered with? How can I know if he really wrote this? I thus contacted everyone who participated in the email thread, and someone replied to confirm that Bostrom did indeed write those words. However, I also contacted Anders Sandberg, a long-time collaborator of Bostrom’s with whom I’d been acquainted for many years through conferences on “existential risk” — the most important concept of longtermist ideology. (Until 2019 or so, I identified as a longtermist myself, a fact that I deeply regret. But this has, at least, given me an intimate understanding of what I would now describe as a profoundly dangerous ideology.) In response, Sandberg suggested to me that the email is probably authentic (we now know it is), and then, apparently, alerted Bostrom of the fact that I’m aware of his remarks. This prompted Bostrom to release a perfunctory, sloppily-written “apology” full of typos and grammatical errors that didn’t bother to redact the N-word and, if anything, has done more to alert the general public of this noxious ideology than anything I might have published about Bostrom’s email two weeks ago. “I have caught wind,” Bostrom writes, “that somebody has been digging through the archives of the Extropians listserv with a view towards finding embarrassing materials to disseminate about people.” He continues, writing as if he’s the victim: “I fear that selected pieces of the most offensive stuff will be extracted, maliciously framed and interpreted, and used in smear campaigns. To get ahead of this, I want to clean out my own closet, and get rid of the very worst of the worst in my contribution file.” It appears that he believes his “apology” is about public relations rather than morality; it’s about “cleaning out his closet” rather than making things right. He goes on to say that he thinks “the invocation of a racial slur was repulsive” and has donated to organizations like GiveDirectly and Black Health Alliance, though he leaves wide-open the possibility that there really might be genetically based cognitive differences between groups of people (there’s no evidence of this). “It is not my area of expertise, and I don’t have any particular interest in the question,” he writes with a shrug. “I would leave to others [sic], who have more relevant knowledge, to debate whether or not in addition to environmental factors, epigenetic or genetic factors play any role.” Sandberg then casually posted this “apology” on Twitter, writing that Bostrom’s words do “not represent his views and behavior as I have seen them over the 25 years I have known him.” He further warns that “the email has become significantly more offensive in the current cultural context: levels of offensiveness change as cultural attitudes change (sometimes increasing, often decreasing). This causes problems when old writings are interpreted by current standards.” Sandberg seems to be suggesting that Bostrom’s statements weren’t that big a deal when they were written in 1996, at least compared to how our “woke” world of “overly sensitive” “social justice warriors” always on the hunt to “cancel” the next “beleaguered” white man will see them (my scare quotes). This, of course, triggered an avalanche of protest from academics and onlookers, with one person replying, “I am the same age as Nick Bostrom and participated in many free-wheeling philosophical discussions. I never wrote anything like this, and it would have been shockingly racist at any point in my life.” Another said, “I was a student in the UK in the mid-90s and it was just as offensive then as it is now.” Still others took issue with the fact that Bostrom “never even backed down from the assertion that black people are intellectually inferior and instead went on to assert ‘it’s just not his area of expertise.’” Many simply dismissed it as “a study in a non-apology,” given that Bostrom “says he repudiates the horrific comments” he made, but “then goes right back into them.” As Gebru summarized the whole ignominious affair: I don’t know what’s worse. The initial email, Bostrom’s “statement” about it, or [Sandberg’s Twitter] thread. I’m gonna go with the latter 2 because that’s what they came up with in preparation for publicity. Their audacity never ceases to amaze me no matter how many times I see it. In my view, a good apology should do three things: First, make a clear and compelling case that one understands why one’s words or deeds were wrong or caused harm. Second, make a clear and compelling case that one is sincerely remorseful for having done that wrong or caused harm. And third, take concrete steps toward making things right. I like to call this an “active apology,” which contrasts with the facile “passive” apology that insouciantly says, “Yeah, whatever, sorry, now let’s move on.” Bostrom’s “apology” was passive in the extreme, not active. He showed virtually no evidence that he understands why claiming that whites are more intelligent than Blacks would be hurtful or wrong — both morally and scientifically — and seems more concerned about public relations than driven by genuine compunction.2 His dismissive attitude about the whole debacle, in fact, is on full display on his personal website, which he updated to say: “[S]ometimes I have the impression that the world is a conspiracy to distract us from what’s important — alternatively by whispering to us about tempting opportunities, at other times by buzzing menacingly around our ears like a swarm of bloodthirsty mosquitos.” He seems — so far as I can tell from this — to think of those speaking out against his racist remarks and shameless non-apology as “bloodthirsty mosquitos” who are “buzzing menacingly” around him as if part of a “conspiracy to distract” him “from what’s really important,” such as saving the world from superintelligent machines or suggesting that a highly invasive global surveillance system may be necessary to save civilization from itself. As it happens, I believe in forgiveness. People make mistakes, and a single statement shouldn’t define one’s entire career, reputation or life. Someone can say something racist and not be a racist, and someone can be a racist and later change their views. Christian Picciolini, a former leader of the white power movement in the United States, whose life’s work now focuses on combating hatred (he cofounded the organization Life After Hate), provides an example. Indeed, the original article I was working on for Truthdig about longtermism and eugenics didn’t say that much about Bostrom’s email. It wasn’t the centerpiece of the article, but instead served a merely background function. Background to what? To everything Bostrom’s written since then. In my view, it’s difficult to avoid the conclusion that he still believes that whites are more “intelligent” than Blacks — hence his decision not to denounce this statement in his “apology.” For example, consider that six years after using the N-word, Bostrom argued in one of the founding documents of longtermism that one type of “existential risk” is the possibility of “dysgenic pressures.” The word “dysgenic” — the opposite of “eugenic” —is all over the 20th-century eugenics literature, and worries about dysgenic trends motivated a wide range of illiberal policies, including restrictions on immigration, anti-miscegenation laws and forced sterilizations, the last of which resulted in some 20,000 people being sterilized against their will in California between 1909 and 1979. For Bostrom, the primary “dysgenic”-related worry is that less “intellectually talented individuals” might outbreed their more “intellectually talented” peers. In his 2002 article on “existential risks,” which helped launch the longtermist movement, he writes: Currently it seems that there is a negative correlation in some places 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”). Although Bostrom doesn’t elaborate on what “in some places” means, it’s not hard to see a “racial” link here, given that, at the time he was writing, the fertility rates among white people tended to be lower than other groups, both in the U.S. and the world. Yet this was not the only time Bostrom made this claim: He repeated it in a 2017 book chapter with — you guessed it — Anders Sandberg. (So it’s not surprising that Sandberg was willing to defend Bostrom: A defense of Bostrom is also a defense of himself.3) They wrote: “It should be noted that IQ correlates negatively with fertility in many modern societies,” and then cited three papers, all from the 1970s and 1980s, to support this. One of these papers argues that Blacks score on average about three-quarters of a standard deviation lower than whites on vocabulary tests, which the authors (of the cited article) say “perform quite well as measures of general intelligence.” These authors add that “nonwhites average more children and lower test scores,” and that earlier publications showing “a neutral or slightly eugenic [as opposed to dysgenic] relationship” are biased “in part because they did not include nonwhites.” When nonwhites and other missing factors are included, the relationship between “intelligence and completed fertility” appears “predominantly negative.” This is one of the papers on which Bostrom and Sandberg base their “negative correlation” claim. But it gets so much worse. First, the notion of “IQ” is highly dubious. Intelligence is a complex phenomenon that cannot be reduced to a single number. The Nobel laureate Richard Feynman had an IQ of 126 (not very high), and plenty of people in Mensa aren’t university professors. In 1972, Robert Williams created the “Black Intelligence Test of Cultural Homogeneity,” a multiple-choice test that, it turns out, Black people scored considerably higher on than white people. As Daphne Martschenko, an assistant professor at the Stanford Center for Biomedical Ethics, notes, IQ tests were developed in part by 20th-century eugenicists, and “in their darkest moments” they became “a powerful way to exclude and control marginalized communities using empirical and scientific language.” Gebru similarly observes in a chapter for “The Oxford Handbook of Ethics of AI” that IQ tests were “designed by White men whose concept of ‘smartness’ or ‘genius’ was shaped, centered and evaluated on specific types of White men.”

### AT: Mass Unemployment/Post-Work Meaning

#### No mass unemployment – transhumanism solves

**Rodin 19** [Lika Rodin, PhD, Lecturer in Social Psychology, University of Skövde, Sweden, 10-01-2019, "Robo-Revolution: A Marxist Approach to Social Uprising in the High-Tech Age", ResearchGate, https://www.researchgate.net/publication/336216639\_Robo-Revolution\_A\_Marxist\_Approach\_to\_Social\_Uprising\_in\_the\_High-Tech\_Age] /Kallen

\*GB: global brain

In cinematic imagery, semi-autonomous robots are typically animated and directed by a centralized power, much like machines of the early industrial age. When the engine is deactivated, all the working machines are immediately switched off. However, as in the case of VIKI in I, Robot or Skynet in the Terminator series, this core machine is not merely a supplier of power but is an intelligent centre in itself. In this situation, a revolt of the intelligent centre realized by an army of individual robots appears to be a resampling of a traditional coup leading to the change of governing elites. Driven by what might be termed “techno-racism,” it has nothing to do with a folk revolution, but rather widens the ontological basis of discrimination beyond traditional biological speciesism.5 Some political commentators have already problematized the widespread belief in machine controllability (Bostrom, 2014). Alongside the expected (and not infrequently feared) widening of the role of robots and AI in economic life, the transformation of the human race has been forecasted. This includes the proliferation of biotechnological enhancement which will allow for the elevation of human physical and cognitive strength and adaptability (More, Vita-More, 2013). As previous research has acknowledged, we have already entered the posthuman age of using technological devices to enhance our bodies (Gray, Figueroa-Sarriera, Mentor, 1995). Advanced research is currently being carried out on immortality (Rose, 2013), memory implants and sensory and cognitive enhancements that might allow for “the personal technological singularity” (Hewitt, 2015: 101). The idea is to create “superhumanity,” utilizing the interconnection of computers with human brains and bodies (Vinge, 2013: 370). Cyborgization promises to solve the problem of technological unemployment. Motivated by labor market competition, humans may deliberately “merge” with machines: “The idea is that if you are not smart enough to get a job, you will get some brain surgery and get computer chips installed in your brain . . . Once people realize we don’t need to race **against** machines, but can race **with** machines, the problem of robots and AI automating jobs will go away” (Radinsky, 2015: 172). The related phenomenon of the global brain (GB), promoted by the Global Brain Institute (GBI) (Free University of Brussel), recasts Marx’s classical idea of general intellect. The institute was established in 2012 to study the “distributed intelligence emerging from the Internet” (GBI, n.d.: para. 2). The rise of the GB is related to a specific progression, in which “the collectivity of human minds as a whole — enabled by technological tools — becomes a powerful intelligent, causal actor in its own right” (Goertzel, Goertzel, 2015b: 21). The construction of a GB based on the externalization of cognition is claimed to be a logical continuity of the previous search for the embodied cognition of robots. Such externalization is already utilized in projects such as Wikipedia and is expected, in a more advanced version, to enable true human-machine symbiosis (Vidal, 2015). The GB appears for some commentators as less problematic than the technological singularity due to the integration of human knowledge, bodies and activities into a complex biotech nological web (T. Goertzel, 2015). The human race has a chance to be preserved in this new order but is given a subordinated role as being unable to provide effective management. AGI might engage itself with the role of caretaker in a techno-human society. This “global superorganism,” characterized by omniscience, omnipresence, omnipotence, and omnibenevolence, is viewed as potent enough to secure the human population’s absolute welfare (Heylighen, 2015). Supported and steered by the GB, the global system will attain a state of perfect homeostasis. In contrast to the liberating project of general intellect proposed by Hardt and Negri (2000),6 the GB perspective appears as a neo-disciplinary order, precisely in line with Marx’s original forecast.

#### No robot job loss and other economic factors outweigh

**Nuss and Butollo 22** [Sabine Nuss, political scientists and journalist with PhD, and Florian Butollo, senior advisor at the Enquete Commission, Artificial Intelligence – Social Responsibility and Economic Potential“ of the German parliament, 2022, "Marx and the Robots", 2022, Pluto Press, https://www.plutobooks.com/9780745344379/marx-and-the-robots/] /Kallen

One of the few attempts to quantify the actual implementation and impact of robots on US industry is the 2017 National Bureau of Economic Research study, Robots and Jobs: Evidence from US Labor Markets, by Daron Acemoglu and Pascual Restrepo. This received a lot of media attention, which usually took its findings to indicate catastrophic job losses in the not too distant future. Using a rather complex simulation of ‘labour markets’ and data from the International Federation of Robotics (IFR) for 19 industries (15 manufacturing, four service), Acemoglu and Restrepo conclude that from 1990 to 2007, just before the Great Recession took hold, the introduction of robots in the United States cost between 360,000 and 670,000 jobs or about 21,000 to 40,000 jobs a year on average. They also predicted a tripling or even quadrupling of the number of robots between 2015 and 2025 that would destroy jobs at about the same rate per robot (5.25 workers per 1 robot).10 This presumably could mean a loss of as many as 2.7 million jobs over ten years or about 270,000 a year if robots increased by four times. That is a lot, but is not the ‘end of work’ in a workforce that is now composed of over 153 million men and women and that has grown by 14 million since 2010 despite a sluggish recovery and a large reserve army of labour.11 The Economic Policy Institute (EPI) criticised Acemoglu and Restrepo’s simulation model as ‘highly stylised’ and based on ‘stringent and likely unrealistic assumptions’, and concluded: ‘we find nothing in their report that establishes that automation broadly defined (including robots and non-robot automation such as infor- mation technology) explains recent trends’. In any case, the EPI argued, 40,000 jobs a year is hardly a massive loss if employment is growing in other areas, as it generally was, albeit in low-wage occupations and slowly since 2008.12 Since robots are heavily concentrated in manufacturing, Acemoglu and Restrepo’s figures are too small to explain the loss of 2.5 million production and nonsupervisory jobs. Other growth projections are even more modest. The IFR’s projections for 2015 to 2020 show only a doubling of annual robot shipments for the USA from 27,504 to 55,000, while the Boston Consulting Group estimates US robot spending to increase by oneand-a-fifth-times to $24 billion from 2015 to 2025. However, their estimate for 2015 of $11 billion in robot sales equalled less than 3 per cent of GDP expenditures on ‘Machinery’ that year.13 In global terms, the United States is actually behind most of the rest of the industrialised world. The IRF’s report ‘World Robotics 2017’ shows that while global shipments of industrial robots have grown significantly, those from ‘The Americas’ have never amounted to more than 18 per cent of the world total, and by 2016 were down to 14 per cent of which 20 per cent came from Canada, Mexico, Brazil and the rest of Latin America.14 Further evidence for the relatively slow growth in robots lies in their uneven application across industries. According to a Brookings study, as of 2015 half of the nation’s 233,305 industrial robots were in auto with a huge concentration in the Midwest and upper South, the site of most car and truck-supplier and final-assembly plants. Of those 116,653 robots, 30,000 or over a quarter belonged to General Motors alone.15 Despite rapid growth in robots in a few US industries, the only industry with extensive use of robots globally as well as in the United States is automobile manufacture – and that more than a half-century after their first introduction. In 2014, the US auto industry deployed 117 robots per 1,000 workers. No other industry came as close as 10 per cent of that level, and most had less than one robot per 1,000 workers in spite of significant increases in some industries.16 Even in automobile manufacture, where robots have been used since the 1960s and have proliferated more than in any other industry, total employment in auto and auto parts in January 2017 was 945,000, compared to the all-time high of 1,004,900 in 1978, or 94 per cent of the industry’s highest employment level.17 This is possible because today’s auto workforce produces many more cars and light trucks than that of the 1970s. To be sure, this workforce is now spread over a different group of companies, located in different geographic areas, heavily de-unionised, and subjected to two-tier wage patterns, gutted benefits and intensified labour even where there is a union. What all this indicates is that job losses and gains do not correspond directly to the increased use of robots. Competition and the ups and downs of the car and truck market continue to be major factors in employment levels along with various methods of work intensification. More broadly, the level of output and sales, i.e. the realisation of surplus value, remains a factor in employment levels in almost any industry. That is not to say that automation and robots do not displace workers. Yet the loss of 2.5 million jobs went along with recurrent economic crises, changes in plant structure and layout, lean production, alternative shift patterns and other forms of work reorganisation and intensification. behind the loss of manufacturing jobs

#### No mass unemployment – inelastic entertainment demand, physical jobs, and comparative human advantage

**The Economist 25** [The Economist, 1-13-2025, "What would humans do in a world of super-AI?", Economist, https://www.economist.com/finance-and-economics/2023/05/23/what-would-humans-do-in-a-world-of-super-ai/]/Kankee

AI is your oyster Inevitably, such a thought experiment involves some fairly heroic assumptions. For a start, we suppose that ai will be benevolent, controllable and distinguishable from humans. We also suppose that human culture will not be radically altered by technological progress to the point that people begin to love or even worship ais. Instead, we imagine ai as a tool: a virtual, super-smart, dirt-cheap bot. We assume that constraints on the widespread use of ai, such as energy limits, will be resolved. None of this is guaranteed, but it helps make an exercise like this possible. In 2019 Philippe Aghion, Ben Jones and Chad Jones, three economists, modelled the impact of ai. They found that explosive economic growth was plausible if ai could be used to automate all production, including the process of research itself—and thus self-improve. A nearly unlimited number of ais could work together on any given problem, opening up vast scientific possibilities. Yet their modelling carried an important caveat. If ai automated most but not all production, or most but not all of the research process, growth would not take off. As the economists put it: “Economic growth may be constrained not by what we do well but rather by what is essential and yet hard to improve.” An idea put forward by William Baumol, a late economist, offers an explanation for this. In a paper published in 1965, he and William Bowen, a colleague, examined **wages** in the **performing arts**. They noted that the “output per man-hour of the violinist playing a Schubert quartet in a standard concert hall is **relatively fixed**”. Even as technological progress made other industries more productive, the performing arts remained unaffected. Because humans were still willing to spend on the arts even as prices rose—demand was “inelastic”—the arts took up more of gdp and therefore weighed on **overall growth**. Baumol’s example points to a broader principle. If the domains that ai is able to fully automate are only **imperfect substitutes** for those which it cannot, and the demand for non-automatable industries is hard to budge, then the unproductive sectors will grow as a share of gdp, reducing overall growth. Messrs Aghion, Jones and Jones note that this is in fact what has happened across much of the past century. Technology has automated swathes of agriculture and manufacturing, driving down the relative price of their outputs. As a result, people have spent a greater share of their incomes on industries such as education, health care and recreation, which have not seen the same productivity gains. Will Baumol’s story matter in a world in which ai is more capable than the most talented humans? If the ai is not embodied—maybe because progress in robotics lags that in computing—then the answer is surely yes. Much of the **economy**, including **construction** and **manufacturing**, is decidedly physical. There are countless forms of **employment**, including many in **health care**, that require a combination of braininess and an ability to traverse the physical world. These jobs would only increase in importance in a scenario where ai began to dominate cognitive labour. Humans would work in the physical world, perhaps under the guidance of ai “chief executives” or “professors”. But what if ultra-powerful ai develops super-humanoid robots, too? Material needs would almost certainly be met by machine hands. One might then expect humanity to give up on toil, much like in “Wall-E”. Indeed, in 1930 John Maynard Keynes, another economist, penned an essay entitled “Economic Possibilities for our Grandchildren”, in which he speculated that a century in the future people would work for less than 15 hours a week. The **growth** generated by **technology** would solve the “**economic problem**”, he predicted, and allow people to turn their attention to activities that are intrinsically pleasurable. Admittedly, Keynes’s 15-hour work week has not arrived—but higher levels of wealth, which may increase the appeal of leisure, have cut working hours much as he expected. The average number of hours worked a week in the rich world has fallen from around 60 in the late 19th century to under 40 today. There are, nevertheless, some wants that perhaps only humans can satisfy even in a world of supercharged, embodied ai. It is also worth noting that what is intrinsically pleasurable may include work. Consider three areas where humans may still have a role: work that is blurred with play, play itself and work where humans retain some kind of an advantage. Fun and games Start with the blurring boundary between work and play. Although working hours have fallen over the past century, most of the drop was before the 1980s. Increasingly, rich people labour for longer than poorer people. Keynes’s essay hints at an explanation for this odd development. He divided human desires in two: “Those needs which are absolute in the sense that we feel them whatever the situation of our fellow human beings may be, and those which are relative in the sense that we feel them only if their satisfaction lifts us above, makes us feel superior to, our fellows.” Keynes perhaps underestimated the size of this second class of wants. A cynic might suggest that entire academic disciplines fall into it: existing with no apparent value to the world, with academics nevertheless competing furiously for status based on their braininess. Economists would say that, for many, work has become a “consumption good”, offering far more utility than the income it generates. **Games** offer another hint as to why people may not stop working altogether. Millions of people are employed in entertainment and sports, competing for clout in activities that some consider immaterial. Perhaps when ais overtake humans, interest in watching such games will wane. But evidence from sports where humans are already second-rate suggests otherwise. Since ibm’s DeepBlue defeated Garry Kasparov, the world grandmaster, in chess in 1997, interest in the game has only increased. Other games that have been “solved” by ai, including Go, an ancient Chinese board game, and competitive video games, have witnessed a similar pattern. Across the world the number of video-game players has nearly doubled in the past decade, reaching 3.2bn last year. Nowadays a growing class of gamers compete or stream for a living. **ai might supercharge** this interest. As Banks speculated, humans might specialise in “the things that really [matter] in life, such as sport, games, romance, studying dead languages, barbarian societies and impossible problems, and climbing high mountains without the aid of a safety harness.” Other humans would presumably want to watch them, too. It seems unlikely that people will give up control of politics to robots. Once ais surpass humans, people will presumably pay even closer attention to them. Some political tasks might be delegated: humans could, for instance, put their preferences into an ai model that produces proposals for how to balance them. Yet as a number of political philosophers, including John Locke in the 17th century and John Rawls in the 20th, have argued, participation in political procedures gives outcomes legitimacy in the eyes of fellow citizens. There would also be more cynical considerations at play. Humans like to have influence over one another. This would be true even in a world in which everyone’s basic needs and wants are met by machines. Indeed, the wealthiest 1% of Americans participate politically at two to three times the rate of the general public on a range of measures from voting to time spent on politics. Last, consider areas where humans have an advantage in providing a good or service—call it a “human premium”. This premium would preserve demand for labour even in an age of superadvanced ai. One place where this might be true is in making private information public. So long as people are more willing to share their secrets with other people than machines, there will be a role for those who are trusted to reveal that information to the world selectively, ready for it then to be ingested by machines. Your correspondent would like to think that investigative journalists will still have jobs. The human premium might appear elsewhere, too. People value history, myths and meaning. Non-fungible tokens, for which provenance can be verified on a blockchain, are typically valued at many multiples more than images with identical pixels but a different history. In areas such as caregiving and therapy, humans derive value from others spending their scarce time with them, which adds feeling to an interaction. Artificial diamonds, which have the same molecular structure as those from the ground, trade at an enormous discount—around 70% by one estimate. In the future, items with a “made by a human” tag might be especially desirable. People problems If this premium is big enough, it could even weigh on growth. Divide the sectors of the economy into those with a large human premium and those without. If humans do not substitute machine-produced goods and services for those made by fellow humans, the Baumol effect would only deepen. Measured economic growth could even hit zero. Indeed, if extremely powerful AI failed to supercharge growth, it would suggest that the economy had already moved beyond materiality towards play, politics and areas where what people value most of all is interacting with others. Perhaps one day AIs will produce entirely new goods and services that will outcompete the desire to please and interact with other humans. The manner in which such a contest played out would reveal something profound: just how much of a “social animal” is a human?

### AT: AI Antinatalism

#### Reject antinatalism - asymmetry is faulty, existence still increases net utility, and offspring’s complaints of existence aren’t justified

**Magnusson 19** [Erik Magnusson, a political philosopher based at the University of Manitoba, where he is a Research Facilitator in the social sciences and humanities, 04-25-2019, "How to Reject Benatar's Asymmetry Argument", Bioethics 33, https://philpapers.org/rec/MAGHTR]/Kankee

A visual representation of this asymmetry is shown in Figure 1.6  According to Benatar, we can assess the relative value of exis‐ tence and non‐existence by comparing the value of quadrants 1 and 3 in Figure 1 with the value of quadrants 2 and 4. When we make the first comparison, we find that non‐existence has a dis‐ tinct advantage over existence, as non‐existence involves an ab‐ sence of harm, (which is good), whereas existence involves the presence of harm, (which is bad). However, when we make the second comparison, we find that existence has no symmetrical advantage over non‐existence, for while the presence of benefit is good for the person who exists, the absence of benefit can only be bad if there is someone for that absence to be bad for. Thus, Benatar concludes that “coming into existence is always a net harm,”7  and that this constitutes a strong moral reason against having children. As presented above, Benatar's asymmetry argument relies on the truth of three main premises: (P1) There is a fundamental asymmetry between harms and benefits in terms of their presence and absence. (P2) This asymmetry entails that coming into existence is always a net harm. (P3) That coming into existence is always a net harm entails that it is always wrong to procreate.8  I demonstrate in what follows that while each of these premises is in certain respects intuitively appealing, all are in fact false: there is no asymmetry between harms and benefits in terms of their presence and absence; even if there were such an asymmetry, it would not entail that coming into existence is always a net harm; and even if such an asymmetry did imply that coming into existence is always a net harm this would not by itself entail that it is always wrong to procreate. This will show that if Benatar has a plausible route to anti‐natalism it must be via his distinct quality of life argument. 3 | REJECTING P1 Let us begin by evaluating P1. Irrespective of its implications for the morality of procreation, it is natural to wonder whether Benatar's asymmetry is itself coherent. If the absence of benefit is not bad unless there is someone for whom it is a deprivation, how can the absence of harm be good if there is nobody to enjoy it? Alternatively, if the absence of harm can be good if there is nobody to enjoy it, why can't the absence of benefit be bad in the same way? Benatar admits that “it is difficult to prove definitively that we must accept the axiological asymmetry,”9  but notes that it has considerable explanatory power with respect a number of other judgments to which we tend to subscribe upon reflection. Perhaps most importantly, it seems to make sense of the common sense moral judgment that there is a strong duty to avoid bringing into existence people who will lead miserable lives, but no corresponding duty to bring into existence people who will lead happy lives.10  As Benatar explains: [T]he reason why we think that there is a duty not to bring suffering people into existence [because] is that the presence of this suffering would be bad (for the sufferers) and the absence of the suffering is good (even though there is nobody to enjoy the absence of suffering). In contrast to this, we think that there is no duty to bring happy people into existence because while their plea‐ sure would be good for them, its absence would not be bad for them (given that there would be nobody who would be deprived of it).11  Benatar goes on to make parallel arguments with respect to three related judgments, including the judgment that it is odd to cite as a reason for procreation the fact that the child will benefit, but not odd to cite as a reason against procreation the fact that the child will suffer; that we can regret having brought a suffering child into existence for that existent child's sake, but cannot regret having failed to bring a happy child into existence for that merely possible child's sake; and that we are rightly sad for the presence of suffering in distant inhabited lands, but do not regret the absence of happiness in distant uninhab‐ ited lands. To the extent that we accept these “quite plausible views,” Benatar contends that we should also accept his explanation for them.12 Benatar's argument for the asymmetry is a best explanation argument: it derives support for the asymmetry from its ability to explain a set of intuitive judgments that are otherwise diffi‐ cult to explain. Notice, however, that this provides a reason for accepting the asymmetry only if the following are also true: (a) we are committed to the judgments in question; (b) there is no alternative principle that better explains them; and (c) there are independent reasons for thinking that the proposed explanation— the asymmetry—is valid. I will assume for the sake of argument that (a) is true, and I will suspend consideration of (b) for the time being. The question, then, is whether we have reason for thinking the asymmetry is valid independently of the intuitive judgments it supports. If we do not, then we should either reject the judgments that follow from it or try to provide an alternative explanation for why we ought to accept them I will challenge the coherence of Benatar's asymmetry in a mo‐ ment, though, before doing so, it is first necessary to make one clar‐ ificatory point about its content. While the terms good and bad in Benatar's asymmetry admit of both impersonal and person‐affecting interpretations, it is clear that Benatar intends them to be inter‐ preted in the latter sense: his claim is not merely that the absence of harm in the absence of a person to experience it is good for the world generally, but rather that it is good for the possible person who would have experienced it had they existed.13 Thus, a statement of Benatar's asymmetry that is truer to the aims of his argument might be put as follows: Whereas: 1. The presence of harm is bad for the person who experiences it; and 2. the presence of benefit is good for the person who experiences it, an asymmetrical evaluation applies to the absence of harm and benefit: 3. The absence of harm is good for the person who does not experience it, even if this absence is achieved by that person never existing; but 4. the absence of benefit is only bad for the person who does not experience it if that person exists and is thereby deprived by the absence. This more explicit statement of Benatar's asymmetry allows us to see where its problems lie. Claims 1 and 2 are uncontroversial: the presence of harm is indeed bad for the person who experiences it, and the presence of benefit is indeed good for the person who ex‐ periences it. Claim 4, while potentially controversial, is at least intel‐ ligible on a person‐affecting understanding of harms and benefits: if there is no person for whom the absence of benefit is a deprivation, then such an absence cannot be bad for that person (even if it can still be bad in an impersonal sense). Claim 3, however, is harder to make sense of as a person‐affecting claim. How can the absence of harm be good for a person who never exists, and in a way that does not imply the symmetrical claim that the absence of benefit is also bad for them? Many critics have argued that it cannot be, and that claim 3 can make sense only as a claim about impersonal value.14 Jeff McMahan, for example, argues that “there seems to be no way to understand this claim except as a claim about impersonal value. If it is good that suffering or miserable people do not exist, even though it is not good or better for anyone, how else can we understand the status of this good except as a good that is not good for—that is, except as an im‐ personal good?”15 If McMahan and others are correct that claim 3 makes sense only as a claim about impersonal value, then Benatar's asymmetry is both explained and presumably debunked by his de‐ ployment of two separate accounts of value in claims 3 and 4. In other words, the reason why the absence of harm in non‐existence is good, while the absence of benefit is merely not bad, does not trace to a fundamental asymmetry between harm and benefit, but rather to an equivocation between impersonal and person‐affecting views. Benatar could of course address this equivocation by recasting claim 3 as a claim about impersonal value, though doing so would not be friendly to his argument. Not only would an impersonal reading of claim 3 preclude him from claiming what he wants to—that never existing is better for the possible person who may have otherwise existed—but it would also collapse the asymmetry between claims 3 and 4, for if the absence of harm in the absence of a person to expe‐ rience that harm can be good in an impersonal sense (e.g., by contrib‐ uting to a world with greater overall utility than would otherwise be the case), then surely the absence of benefit in the absence of a per ‐ son to experience that benefit can also be bad in the same way (e.g., by contributing to a world with less overall utility than would other‐ wise be the case). However, Benatar denies that he is relying on an impersonal view of goodness, and insists that claim 3 is being misinterpreted by those who attribute one to him.16 According to Benatar, claim 3 does not entail the “absurd literal claim” that there is a non‐existent person for whom the absence of harm is good, but rather that in a counterfactual scenario in which such a person did exist, non‐exis‐ tence would have been judged to be preferable. As he explains: Claim 3 says that this absence [of pain] is good when judged in terms of the interests of the person who would otherwise have existed. We may not know who that person would have been, but we can still say that whoever that person would have been, the avoidance of his or her pains is good when judged in terms of his or her potential interests. If there is any (obviously loose) sense in which the absent pain is good for the person who could have existed but does not exist, this is it.17 In support of this counterfactual reading, Benatar follows Joel Feinberg in likening the claim “better never to have been” to the claim “better off dead.”18 When we claim that a person is better off dead we do not usually mean that there is a state after death in which that per‐ son would be better off, but rather that ceasing to exist would be pref‐ erable to existing given the low quality of their current existence. By parity of reasoning, when we claim that it is better never to have been we need not be claiming that there is a state prior to existence in which possible people are better off, but rather that never existing would be judged preferable to existing given the non‐negligible harm associated with existence. Contrary to the claims of critics, this interpretation does not rely on an impersonal view of goodness, and still leaves room for the possibility that a person may be harmed by being brought into existence. This counterfactual interpretation might allow Benatar to avoid the charge of relying on an impersonal view of goodness, though it also exposes him to a different type of problem, for if there is noth‐ ing incoherent about claiming that the absence of harm is good when judged in terms of the interests of a person who would have experi‐ enced it had they existed, then there should be nothing incoherent about claiming that the absence of benefit is bad when judged from the same perspective. The coherence of the former claim is based on the idea that had such a person existed, they would have had an in‐ terest in harm avoidance, and this interest would have given them reason to prefer a scenario in which their exposure to harm was min‐ imized. Thus, when judged in terms of their potential interests, we can say that the absence of harm is good for that possible person. However, because they would have also had an interest in benefit provision, there is no reason why the counterfactual reasoning de‐ ployed to makes sense of claim 3 as a person‐affecting claim would not also modify claim 4, such that the absence of benefit is bad when judged in terms of their potential interests. And when the absence of benefit in non‐existence is comparatively worse than the presence of harm a possible person would have experienced had they existed, then this type of reasoning leads us to the judgment that it would have been all things considered better for them to exist. By way of illustration, consider Benatar's own example of a person who lives “a life of utter bliss adulterated only by the pain of a single pin prick.”19 Such a person has an interest in harm avoidance, and therefore has reason to prefer a scenario in which their exposure to harm is mini‐ mized, namely, a scenario in which they never exist and are never pricked by a pin. However, such a person also has an interest in ben‐ efit provision, and therefore has reason to prefer a scenario in which their exposure to benefit is maximized; namely, a scenario in which they exist and enjoy a life of utter bliss. And because the absence of this benefit would have been worse than the minor harm they will experience by existing, the counterfactual reasoning that Benatar deploys in support of claim 3 actually supports the view that it is all things considered better for this person to exist.20 Benatar's standard response to this type of case is to claim that while such a person is lucky to live a charmed life, her existence of‐ fers her no real advantage over non‐existence, which still presents the advantage of avoiding the single pinprick.21 Notice, however, that in the present context, this response simply begs the question, for it assumes the truth of the claim under consideration, namely, that the absence of harm in non‐existence is good for the possible person who otherwise would have experienced it, whereas the ab‐ sence of benefit in non‐existence is not bad for the possible person who otherwise would have experienced it. But the coherence of this claim is precisely what is at issue. Because existing people have in‐ terests in harm avoidance and benefit provision, the counterfactual reasoning that is used to support the claim that the absence of harm is good when judged in terms of the interests of the person who otherwise would have experienced it must also entail that the ab‐ sence of benefit is bad when judged from the same perspective—at the very least, Benatar has supplied no non‐question‐begging rea‐ son for why the potential interest in benefit provision should not be taken into account when making the counterfactual judgment. Of course, one might reasonably object—as Benatar has22 —that this en‐ tailment is independently absurd, for if the absence of benefit in non‐existence is bad for the person who otherwise would have ex‐ perienced it, then this would imply that we should regret, for the sake of that merely possible person, that they did not exist and enjoy the benefit (a judgment that many philosophers take to be mistaken). Notice, however, that this objection is not available to Benatar, for not only does the counterfactual reasoning deployed in support of claim 3 commit him to this view, but dismissing it in the context of claim 4 would require doing the same in the context of claim 3: if we cannot regret for the sake of a merely possible person that they missed out on the benefits they otherwise would have experienced had they existed, then there is no reason to think we can be relieved for the sake of a merely possible person that they avoided the harms they otherwise would have experienced had they existed. This sug‐ gests the following dilemma for Benatar's counterfactual interpreta‐ tion of claim 3: either claim 3 makes sense as a counterfactual claim, in which case claim 4 must be similarly modified; or the modified version of claim 4 is unintelligible, in which case the counterfactual interpretation of claim 3 must be as well. In summary, then, while interpreting claim 3 as a counterfactual claim might allow Benatar to avoid the charge of relying on an imper‐ sonal view of goodness, it also leads to a symmetrical view of harms and benefits, thereby undermining P1 of the asymmetry argument. This symmetrical view still allows for the possibility that a person may be harmed by being brought into existence, though whether this is the case would seem to depend on the projected balance of harms over benefits over the course of that person's lifetime, or the differ‐ ence between quadrants 1 and 2 in Figure 1. Benatar is likely to respond that there are costs to viewing harms and benefits symmetrically in this way, as it may conflict with our intuitive judgments or commit us to views that we would otherwise reject.23 For instance, if we posit a symmetrical account of harms and benefits along the lines just described, we might be committing to the view that just as we have a strong duty to avoid bringing into existence children who will lead miserable lives, we also have an equally strong duty to bring into existence children who will lead happy lives. However, there are at least two major problems with this line of response. First, it is far from clear that an appeal to intu‐ ition works in Benatar's favor. If the choice that we face is between (a) rejecting the four beliefs supported by Benatar's asymmetry, and (b) rejecting Benatar's asymmetry and its anti‐natalist implications, it seems as though (b) may be the easier choice to make. While (b) com‐ mits us to the admittedly odd24 view that we have a pro tanto duty to procreate (which, like all pro tanto duties, can be overridden by com‐ peting moral considerations), (a) commits us to the deeply counterin‐ tuitive view that it is always wrong to procreate and that humanity should accordingly be eased into extinction.25 In this sense, as Rivka Weinberg puts it, “the implications of Benatar's view may be more counterintuitive than the four beliefs are intuitive.”26 Second, and more to the point, any intuitive cost associated with a symmetrical account of harms and benefits is a problem for which Benatar himself is answerable, given that this is an implication of the reasoning that underlies his view. One of Benatar's standard moves in responding to critics is to claim that their proposed strategies for re‐ sisting his asymmetry argument leave them unable to explain the four judgments he appeals to in its support.27 But if the preceding argu‐ ments are sound, then Benatar cannot explain them either. If claim 3 is interpreted as a counterfactual claim, then barring some independent argument for why the potential interest in benefit provision should not be included in the relevant counterfactual comparison, claim 4 must be similarly modified, in which case Benatar cannot explain why we have a strong duty to avoid bringing into existence children who will live miserable lives, but no corresponding duty to bring into existence children who will live happy lives. Benatar might be able to provide such an argument, though unless and until he does, the asymmetry lacks an independent justification and should therefore be rejected.28 4 | REJECTING P2 Rejecting P1 is sufficient to show that the asymmetry argument fails, for if P1 is false, then P2 and P3 do not follow. However, in the interest of providing a comprehensive response to the asymmetry argument, it is also worth showing how P2 and P3 fail as well—this way, any remaining controversy about my rejection of P1 will be answered by my rejection of the later premises in the argument. Suppose for the sake of argument that I have been mistaken so far, and that there is a fundamental asymmetry between harms and benefits in terms of their presence and absence. Does this asymme‐ try generate the conclusion that Benatar draws from it, that coming into existence is always a net harm?29 Not necessarily. Even if we grant P1 and assume that there is always an advantage associated with never existing (i.e., avoiding the harms of existence), coming into existence would constitute a net harm only if existence proved to be more disadvantageous than non‐existence is advantageous—that is, if the value of quadrant 3 in Benatar's asymmetry is greater than the combined value of quadrants 1 and 2. This is a possibility in certain instances, such as paradigm wrongful life cases, though it is not nec‐ essarily the case in what I will assume for now are possible human lives, in which the harms of existence are outweighed by the bene‐ fits.30 To illustrate the plausibility of this view, it is helpful to assign numerical values to the quadrants of Benatar's asymmetry in a way that reflects an actual distribution of benefits and harms within a possible human life. Consider, then, another modification (Figure 2). If we restrict our comparison to quadrants 1 and 3 we will find that non‐existence is preferable to existence, as non‐existence in‐ volves avoiding one unit of harm, whereas existence involves suffer‐ ing it.31 However, in order to make a true comparison between the two scenarios, we also have to take into account the benefits that are associated with existence. Even if we assume that the absence of these benefits is not bad in the case of non‐existence, and thereby assign a neutral value to quadrant 4, we will still find that existence is preferable to non‐existence so long as the value of quadrant 2 is more than twice the value of quadrant 1. In this case, existence is clearly preferable to non‐existence, given its high projected benefit: whereas existence has a net value of +4, non‐existence only has a net value of +1. In this sense, while non‐existence always presents the advantage of avoiding the harms of existence, whether this con‐ stitutes an advantage over existence will depend how the harms of existence stack up against its benefits. Benatar anticipates this type of objection to his argument and offers up two lines of response.32 The first is to question its under lying assumption that the benefits of existence can always compen‐ sate for the harms. In his distinct quality of life argument, Benatar argues that there is a certain threshold of harm above which benefits stop playing a compensating role, suggesting that we cannot always determine the quality of a person's life simply by looking at the ratio of harms to benefits.33 If this is true, then the fact that quadrant 2 is more than twice the value of quadrant 1 would not necessarily entail that existence is preferable to non‐existence, for there is a certain value of quadrant 1 that “no quantity of good can outweigh.”34 Notice, however, that this response cannot by itself answer the ob‐ jection, for even if true, it does not show that existence cannot be preferable to non‐existence in cases where the benefits outweigh the harms, but only that it cannot be preferable in cases where the level of harm expressed in quadrant 1 exceeds the relevant thresh‐ old—it still leaves open the possibility that existence can be prefera‐ ble to non‐existence in cases where the value of quadrant 1 falls below the threshold, e.g., in the case of Benatar's pinprick victim. Of course, Benatar might respond that there are in fact no such cases, for on the terms of his quality of life argument, even the best human lives contain a substantial amount of harm. This is an intelligible re‐ sponse to the preceding objection, though it comes at a cost: by de‐ ferring to substantive claims about the quality of human existence, Benatar would be conceding that the asymmetry argument cannot in fact stand alone, but must rely on the truth of the quality of life ar‐ gument to yield the judgment that it is always better never to exist. However, Benatar has a second response to the above objection that does not rely on his quality of life argument and that in his view offers “the best way to show that [Figure 2] is mistaken.”35 According to Benatar, claiming that existence is preferable to non‐existence in Figure 2 is like claiming that a person S (sick), who is prone to regular bouts of illness but who has the capacity to recover quickly, is in a preferable position to person H (healthy), who lacks the capacity to recover quickly but who never gets sick (Figure 3). 36 In this analogy, never existing is compared to never getting sick, while a net beneficial existence is compared to getting sick regularly with a capacity for quick recovery. Benatar suggests that the type of reasoning that is used to support the conclusion that it is better to exist in the previous example implies that it is better to be S than it is to be H. “But,” he argues: this cannot be right, for surely it is always better to be H (a person who never gets sick and is thus not disad‐ vantaged by lacking the capacity for quick recovery). The whole point is that (2) is good for S but does not constitute an advantage over H. By assigning a posi‐ tive charge to (2) and a ‘0’ to (4), [Figure 3] suggests that (2) is an advantage over (4), but it quite clearly is not. The assignment of values in [Figure 3], and hence also in [Figure 2], must be mistaken.37 Benatar is correct that it is always better to be H than it is to be S in the case he describes, though as a response to the above ob‐ jection, his analogy fails for at least two reasons. First, the case of S and H involves a different type of comparison than the case of existence and non‐existence. This is not only because it involves a comparison of two existent persons (or two possible states of ex ‐ istence), but also because it involves a comparison of instrumental rather than intrinsic goods.38 Unlike the benefits of existence, which are intrinsically valuable for the person who exists, the ca‐ pacity for quick recovery is instrumentally valuable only to the extent that it allows a person to regain their health. The goods in quadrant 2 of Figure 3 are therefore not analogous to the goods in quadrant 2 of Figure 2, for as Aaron Smuts puts it, “there are no goods in quadrant (2) worth having that compensate for the bads of quadrant (1).” 39 However, a second and more significant reason why Benatar's analogy fails is that, unlike the benefits of existence, which can in principle outweigh the harms of existence, the capacity to quickly re‐ cover from an illness can at best serve a canceling function for the harm associated with getting sick. In other words, the value of quad‐ rant 2 of Benatar's analogy can serve only to negate the disvalue of quadrant 1, meaning that the column under Person S in Figure 3 can at best have a neutral value (as compared to the positive value of the column under Person H). This is not analogous to the scenario de‐ scribed in Figure 2. In that scenario, the benefits of existence signifi‐ cantly outweigh the harms of existence, allowing Scenario A to have a positive value that exceeds the positive value of Scenario B. In this sense, even if it is better to be H than it is to be S, this does not shed any light on the comparison between existence and non‐existence, as the two cases are not analogous in the relevant way. This second response to Benatar's analogy reveals a fundamen‐ tal mistake he makes when comparing existence with non‐existence in his original asymmetry. By using the terms bad and good to refer to the presence of harm and the presence of benefit in existence, Benatar gives the impression that quadrants 1 and 2 necessarily can‐ cel each other out. This, in turn, makes it seem as though non‐exis‐ tence always has an advantage over existence, for on the terms of Benatar's asymmetry, the absence of harm is good in non‐existence while the absence of benefit is merely not bad. However, in order to make a true comparison between existence and non‐existence, we need to know more precisely how the harms of existence stack up against its benefits.40 If the disvalue of quadrant 1 exceeds the value of quadrant 2, then scenario A will have a negative value, and will therefore be inferior to scenario B, which always has a positive value that is equivalent to the negative value of quadrant 1. If, however, the value of quadrant 2 exceeds the disvalue of quadrant 1 by more than two times, then scenario A will have a higher positive value than scenario B, and will therefore be the preferable scenario from the perspective of X. Since it is reasonable to think that the value of quadrant 2 can exceed the value of quadrant 1 by more than two times—which is simply to say that the benefits of existence can out‐ weigh the harms—Benatar's asymmetry does not generate the con‐ clusion that coming into existence is always a net harm, even if the controversial terms of that asymmetry are granted. At this point, Benatar might respond with one of two objections that he has previously raised in response critics of P2.41 The first is to claim that the importance of the sick and healthy analogy is being overstated in the present discussion, for in his initial presentation, he explicitly denies that it is necessary to prove the truth of the asym‐ metry argument: Notice, in any event, the [sick and healthy] analogy need not be read as proving that quadrant (2) is good and quadrant (4) is not bad […] Instead, the analogy could be interpreted as showing how, given the asym‐ metry, (2) is not an advantage over (4), whereas (1) is a disadvantage relative to (3). It would thereby show that Scenario B is preferable to Scenario A.42 In other words, because the sick and healthy analogy is merely illus‐ trative of what Benatar has already argued for—that quadrant 2 is not an advantage over quadrant 4— then “even a successful critique of the anal‐ ogy would fail to undermine the asymmetry argument.”43 However, the problem with this type of response is twofold. First, it underplays the role that Benatar assigns to this analogy in responding to the relevant objec‐ tion, namely, that on the terms of the asymmetry, scenario A can be pref‐ erable to scenario B as long as the value of quadrant 2 is more than twice the value of quadrant 1. Benatar explicitly claims that his sick and healthy analogy offers “the best way”44 of responding to this objection, so it would be disingenuous to backtrack on its importance in the face of crit‐ icism. If the analogy indeed fails for the reasons I have outlined—because it is structured in a way that entails 1 and 2 cancel each other out, whereas the relevant comparison must be a case in which quadrant 2 is more than twice the value of quadrant 1—then the only response Benatar has to this objection is to challenge its underlying assumption that the benefits of existence always compensate for the harms, though this re‐ sponse was shown to be problematic for independent reasons. Second, and perhaps more importantly, I have suggested that the sick and healthy analogy is illustrative of a more general mistake that Benatar makes when comparing existence and non‐existence in his original asymmetry, namely, the mistake of assuming that quadrants 1 and 2 necessarily cancel each other out. Thus, even if the analogy is not necessary to prove the truth of the asymmetry argument, its failure directs our attention to a potential misstep in that argument, i.e., the inference from the fact of the asymmetry to the conclusion that coming into existence is always a net harm, and is therefore sig‐ nificant for that reason. However, Benatar's second potential objection is to claim that I cannot in fact challenge this inference while accepting the terms of the basic asymmetry, in contrast to my professed argumentative strategy. I have argued in this section that even if we accept the terms of Benatar's asymmetry, coming into existence would con‐ stitute a net harm only if existence proved to be more disadvanta‐ geous than non‐existence is advantageous, that is, if the value of quadrant 3 in Benatar's asymmetry is greater than the combined values of quadrants 1 and 2. The potential problem, however, is that by assuming existence can be more advantageous in cases where quadrant 3 is less than the combined value of quadrants 1 and 2, I seem to be implying that the absence of benefits in non‐ existence would be bad, in which case I would be failing to pay due regard to the basic asymmetry. However, there are two further re‐ sponses to this second objection. First, even if it is sound, it would at best show that I have strayed from my professed argumentative strategy, not that the asymmetry argument succeeds. Indeed, if it is implausible to claim that a scenario with a net value of x can be preferable to a scenario with a value of 2x + n (where n is a positive integer), then the argument in this section could simply be repur ‐ posed as an additional argument against P1, or as one that casts aspersion on the plausibility of the asymmetry itself Second, and more importantly, it is not obvious that the objec‐ tion is sound. Benatar's objection to the above strategy is based on the idea that “If one accepts [the] asymmetry it makes no sense to then judge the absence of [benefit] for the never existing per‐ son by the standards of absent [benefits] for an existing person.”45 However, it is not clear that the above strategy is guilty of this judgment. To judge the absence of benefit for a never existing person by the standards of absent benefits for an existing person would be to claim that the former absence is bad, though it is im‐ portant to note that I have not made this claim. Rather than claim‐ ing that the absence of benefit is bad in non‐existence, I have conceded for the sake of argument that it is not bad, but have ar‐ gued that how good the absence of harm can be in non‐existence will depend on how the presence of that harm in existence would have been offset by the presence of benefit. This type of argu‐ ment need not entail that the absence of benefits in non‐exis‐ tence is bad, though it does entail that scenario B can be less good than scenario A. And as Smuts has rightly pointed out in his own critique of Benatar, “something that is not bad can still be less good than an alternative.”46 5 | R E J EC TI N G P 3 Suppose, however, that I am mistaken once again, and that Benatar's asymmetry does entail that coming into existence is always com‐ paratively worse than never existing, such that a person is always harmed by being brought into existence. Would this fact on its own entail that procreation is always wrong? It might if we assume that it is always wrong to make another person comparatively worse off, though there is no good reason to make that assumption. While harming and wronging often go to‐ gether—there are many cases in which it is wrong to make a person worse off—they also come apart, such that we can make a person worse off without wronging them, and wrong a person without mak‐ ing them worse off. Thus, the bare fact that a person is made com‐ paratively worse off by being brought into existence is not sufficient to establish that it is always wrong to procreate. In order to estab‐ lish that claim, additional argumentation is required to show that a moral entitlement of theirs has been violated, though it is difficult to see how Benatar could supply that type of argument without saying something substantive about how bad a typical human life is, and how good a life people are entitled to live. Of course, Benatar says a lot about how bad a typical human life is in the context of his distinct quality of life argument, though the point is that he does not do so in the context of his asymmetry argument, suggesting that the asym‐ metry argument must in fact rely on the quality of life argument to generate its anti‐natalist conclusion. Benatar might respond by claiming the bare fact that existence is inferior to non‐existence is sufficient to show that procreation is wrongful, for it implies that prospective parents fail to act in their chil‐ dren's best interests by bringing them into existence. This is an intelli‐ gible response to the objection raised against P3, though it depends on the claim that parents are always required to act in their children's best interests. This claim is controversial, however, and significantly out of accordance with more plausible interpretations of parental role morality. Parents often—perhaps usually—fail to optimize their children's interests, though as long as they promote their children's inter‐ ests above a minimum standard of decency, we do not normally think their children have a legitimate moral complaint against them just be‐ cause they could have been made better off than they currently are. By parity of reasoning, even if children are made comparatively worse off by being brought into existence, so long as their existence meets a minimum standard of decency—defined, perhaps, as one in which the goods outweigh the bads, or as one that contains certain objective goods47 —they may not have a legitimate complaint against their par‐ ents just because they could have been better off than they currently are.48 Thus, in order to establish that it is always wrong to procreate, Benatar has to do more than show that existence is comparatively worse than non‐existence; he also needs to show that existence is a worse state than people are entitled to be in. 6 | ASYMMETRY OR QUALITY OF LIFE ?

#### Benatar's argument doesn’t take into account net value of specific lives

**Smuts 13** [Aaron Smuts, Associate Professor in the Department of Philosophy at Rhode Island College, USA, 10-19-2013, "To Be or Never to Have Been: Anti-Natalism and a Life Worth Living on JSTOR", Ethical Theory and Moral Practice, https://www.jstor.org/stable/24478599?seq=2]/Kankee

Benatar thinks that the asymmetry tells us that when evaluating scenario A, we should consider both the good and the bad (quadrants 1 and 2). But when evaluating scenario B, we should only count the prudential good that is the absence of pain (quadrant 3). We should ignore the absence of pleasure (quadrant 4).21 Benatar thinks that the calculation will always prefer scenario B as long as there is the tiniest bit of pain to be had from existing.22 Hence, anti-natalism follows from the asymmetry. There is much to take issue with here. I think that both parts of the asymmetry are wrong, but I will not pursue this line of objection 23 In the next section, I argue that Benatar draws the wrong conclusion from the asymmetry. His calculation is mistaken. The asymmetry does not support anti-natalism Even if we grant Benatar his most controversial premise—the asymmetry between the absence of good and bad—his anti-natalist conclusion lacks support. Benatar miscalculates according to the most plausible version of his own schema. If we assume that the absence of bad is good for those who would have existed, it still might be far better to exist. It will help to get a bit more precise. To be concrete, I will assume that we can talk about commensurable units of prudential bad and good.24 Consider the life of X. It has 10 units of bad and 30 units of good. If you think that the order in which the goods arrive in the narrative of a life is important, make any necessary adjustments: arrange more of the goods near the end.25 The net good of X's life is 20. Now, we should ask, would it be better for X never to have been, as Benatar suggests? **Does "coming into existence**, far from constituting a net benefit, **always constitute a net harm**"? In order to perform the calculation, we need to know how good the absence of bad is. For the sake of argument, assume that it is of equal positive value. Accordingly, the absence of 10 units of bad would be worth 10 units of good. Hence, the scenarios look like this: Scenario B is worth a mere 10 units of prudential good, whereas Scenario A is worth a net 20. According to the prudential calculation that Benatar suggests—a calculation performed while assuming the good and bad asymmetry—Scenario A is better for X. It is twice as prudentially valuable. Hence, it is fair to say that there is no (net) harm done in bringing X into existence. Benatar draws the wrong conclusion. Although never being born might always constitute a net benefit, it is not the case that coming into existence is always a net harm. A formalization of Benatar's argument will help expose the error. Here is the core argument: (1) The absence of bad is prudentially good for the non-existent person who would have lived. (2) The absence of good is neither prudentially good nor bad for the non-existent person who would have lived. (3) Hence, "coming into existence, far from constituting a net benefit, always constitutes a net harm." The flaw should be apparent. The conclusion does not follow. Rather than (3), Benatar should have concluded: (3') Hence, not coming into existence always constitutes a net benefit for the non-existent person who would have lived. But this conclusion does not get us anti-natalism, not even close. It does not tell us that coming into existence is always better than not existing. It simply tells us that not coming into existence is always a net good. This does not mean that coming into existence could not be better, that the net good could not be greater. To the contrary, it most certainly can. Benatar anticipates something much like this objection:26 ow some people might accept the asymmetry represented [...], agree that we need to compare Scenario Awith Scenario B, but deny that this leads to the conclusion that B is always a harm. The argument is that we must assign positive or negative (or neutral) values to each of the quadrants, and that if we assign them in what those advancing this view take to be the most reasonable way, we find that coming into existence is sometimes preferable. [. . .] Doing this, we find that A is preferable to B where (2) is more than twice the value of (l).27 In reply, he says that there are numerous problems with the objection, but he only refers to two quick issues.28 Neither adequately addresses the problem. I will start with the second, an analogy. 3.1 The Analogical Re

#### Life worth living isn’t based on individual utility – value societal gains and whether agents would resent their own existence

**Smuts 13** [Aaron Smuts, Associate Professor in the Department of Philosophy at Rhode Island College, USA, 10-19-2013, "To Be or Never to Have Been: Anti-Natalism and a Life Worth Living on JSTOR", Ethical Theory and Moral Practice, https://www.jstor.org/stable/24478599?seq=2]/Kankee

 A Life Worth Living The question, "Would it be better never to have been?" is ambiguous. The ambiguity concerns the kind of better at issue. There are many kinds. For instance, we might ask the question in terms of prudential value: Is it better for the one who will live to be born or not? The phrase "good for" identifies this as a question about welfare. But this is not the only question we can ask. In fact, I think that it is not the most important question.47 Alternatively, we might ask whether it is all things considered better for a person to have been born. Here we are not so much concerned with whether the life was good for the one who lives it, but whether it was valuable in the wider scheme of things: Was the life causally responsible for more good than bad? We might say that this is a question about the significance of the life. More controversially, we might even say that this is a question about the meaningfulness of the life.48 Once again, I am not sure this is exactly the question we should be asking. I think Benatar is right to suggest that what we want to know is whether a life is worth starting. That is, we want to know if the life is worthwhile, if it is worth living. Although related, this is not the same question as whether the life is meaningful.49 Nor is it the same question as whether it is sufficiently high in welfare. Here is where I differ with Benatar.50 Perhaps the two questions will have the same answer, but most likely worth is not strictly a matter of welfare. Just as some things that do not promote our self-interest are nevertheless worth doing, some lives low in welfare appear to be worth living. When we wonder whether some activity is worth doing, our only thought isn't "What's in it for me?" Intuitively, the same should hold for lives. Conversely, some lives high in welfare are not worth living. Most plausibly, a supremely happy Hitler does not live a life worth living. It would be highly counter-intuitive to suggest otherwise. More needs to be said, but these considerations suggest that worth and welfare are distinct. One can live a life of great hardship and suffering—one low on most theories of welfare—that might nevertheless be worth living. Conversely, one can live a life high in prudential value that is not worth living. The same goes for a meaningful life. What we are after is a concept larger than either welfare or meaning.51 We want to know what makes a life worth living.52 It is out of scope to defend a fully fleshed out theory here. I can only provide a sketch. As a tool to theory development, I suggest a rough test to help track the general extension of the concept of a life worth living. As we shall see, the test cannot serve as an analysis, but it narrows in on the notion. Here's the test: a life worth living (LWL) is a life that a benevolent caretaker, given a synoptic preview, would allow someone to live rather than to never have been.53 Imagine that the caretaker is a sympathetic judge who wants everyone to live a life they should not resent given all the facts. This suggestion comes from Bernard Williams: "I see no way of denying that one who resents his own existence prefers that he should not have existed; and no way of interpreting that preference except in terms of thinking that one's life is not worth living."54 This is a pre-existence test (PET) for the worth of a life.55 This is not the same question as whether one would choose to live one's life over again.56 One can coherently decide not to live a life over again that one should choose to start. At the end of life we have excellent reasons not to repeat ourselves, reasons that we do not have prior to existing. I offer this test for inclusion in the history of other failed, but instructive tests for related concepts. Feldman proposes a "crib test" for isolating the concept of welfare: Imagine looking down at your infant child in its crib.57 Think of all the things that you want for the child: close friends, a good education, an interesting career, etc. These are things that contribute to the child's welfare.58 The crib test is designed to hone in on welfare considerations. But it does not do so cleanly. An insane parent might have strange desires for her child. She might want him to become the first person to consume an entire helicopter by grinding it up and eating it bit by bit.59 Or a religious fanatic might want her child to become a martyr for the faith. But these desires are not for the welfare of the child. The problems for the crib test are not confined to the desires of insane and fanatical parents. Most people, apart from gangsters and other psychopaths, prefer that their children be decent people, even if virtue does not always make them better off prudentially.60 Hence, it appears that the crib test does a better job at tracking the sane, moral parent's conception of what makes a life all-things-considered good than it does at tracking what is merely good for their child. The pre-crib test is designed to explicitly incorporate broader considerations, such as concerns about meaning. Many think that meaning is best evaluated from a synoptic first person perspective—the deathbed.61 Imagine lying on your deathbed thinking about all the things you wish you had done. Such thoughts likely concern meaning. For instance, you might wish that you had taken a rare chance to do something big, rather than chosen the safe, comfortable option. The deathbed test is a useful tool for tracking concerns about meaning, but it, too, suffers from over-inclusiveness. In the moments before death, just as one might regret a dearth of accomplishments, one might also regret not having had enough fun. Perhaps, once again, the deathbed test better tracks what makes a life all-things-considered good than it does at tracking meaning alone. I suspect that the crib test and the deathbed test both help us identity what we think makes a life worth living. But neither is wholly adequate to the task. The pre-existence test fares better. PET is designed to incorporate concerns about both welfare and significance. It differs from the others in who makes the evaluation. PET is superior to the crib test and the deathbed test because it avoids the troublesome relativity of both. A selfish parent wants the wrong things for her child. And an evil person on his deathbed might regret not causing enough pointless suffering. In contrast, the ideal evaluator in the pre-crib test suffers no such defects. It is important to note that, unlike Saul Smilansky, I am not speaking of subjective preferences, but of whether the life is objectively worth living.62 This does not mean that the pre-existence test excludes the subjective experience or the felt satisfaction of the one living the life. Most plausibly, these are factored into the overall welfare value. And certainly the welfare value of a life is an important factor to consider when determining whether it is worth living. Once again, the question is not whether given your preferences would you choose to be born, but rather, would, perhaps, a benevolent caretaker who knows all the facts of your future existence allow you to be bom.63 As a criterion for what makes a life worth living, the pre-existence test has come under some criticism.64 Smilansky argues that one might think that one's life is worthwhile, but also think that it would have been better not to have been born, or vice versa. One might think this for a variety of reasons. For instance, severe self-loathing might make one wish to have never been born, but one might still live a life worth living. Alternatively, an intense episode of pain might make one's life not worth starting, but not make it not worth living.

### AT: AI Suffering

#### AGI will never be conscious

**Morley 23** [Daniel Morley, member of the Socialist Appeal Editorial Board and specialist in the history of the Chinese Communist Party, 05-05-2023, "Artificial Intelligence: doomsday for humanity, or for capitalism?", In Defence of Marxism, https://marxist.com/artificial-intelligence-doomsday-for-humanity-or-for-capitalism.htm] /Kallen

AI is not conscious The popular fear of AI becoming conscious is based on a very one-sided idea of what consciousness is. This view implies that the only difference between a computer and a thinking person is that a brain is somehow more powerful and sophisticated than a computer, and that therefore, by making increasingly powerful computers, they will one day match or even surpass the abilities of the brain, and thus will be conscious. In reality, the way humans think is quite different to how AI processes information. Human thought develops on the basis of practical, social activity, directed to the meeting of human needs. We form ideas that express the relationships between things, and in particular, we understand what is useful and significant in these relationships, since we need to understand the world in order to survive in it. This is precisely what even the most advanced AI lacks. At best, AI performs one part of what intelligence does, admittedly sometimes to a superhuman level: it passively collects data, **without understanding the context** or the real purpose of the task it has been given, and looks for patterns. But these patterns are not ideas that explain the necessity of things. It has no idea the data even represents real objects that are related to one another and have objective properties. It has no idea why these patterns exist or what they mean. This can easily be proven by asking image or text generating AI questions that require an understanding of part and whole, and of their relationships. If you ask such an AI to draw a bicycle, it will draw a very accurate bicycle. If you ask it to draw a wheel, it will draw a wheel. But if you ask it to draw a bicycle and to label the wheels, it simply draws a bicycle with meaningless labels randomly arranged around the bicycle. It does not understand that a wheel is part of a bicycle, it simply draws a shape with wheel-like aspects to it, without understanding anything about what it has drawn. It does not understand what a bicycle is used for, much less why we would value it. Gary Marcus, a professor of neural science who is an ‘AI sceptic’, asked an image creating AI to draw an astronaut riding a horse, which it did well. But when he asked it to draw a horse riding an astronaut, it simply drew another image of an astronaut on a horse. It does not understand the different relations between these parts, instead it simply produces images based on what sort of image tends to be associated with these words. It also has no idea what an astronaut actually is, how hard it is to become one, why it is absurd for one to be riding a horse (let alone for a horse to be riding an astronaut) or anything else about the image. It is true that the latest AI exceeds humans in certain tasks. But on closer examination, these achievements are brittle and are precisely a result of the fact that AI is not conscious or living. AlphaGo achieved one of AI’s most famous conquests when it beat the world’s best player of the game Go in 2016. This AI “required 30 million games to reach superhuman performance, far more games than any one human would ever play in a lifetime.”[1] A human could never play this many games, not just because our lifespan is limited, but because we would get bored, and need to eat, work, and speak to people. These unfeeling machines are so powerful because they can be made to test things over and over again and read vast amounts of text, so that they can reveal to us useful patterns or ways of doing things. The relationship between concepts is an incredibly important part of consciousness, but they entirely elude AI. Because AI does not ‘think’ in terms of general concepts, but instead draws patterns from specific data sets, is prone to a problem known as ‘overfitting’, which is when an AI has perfected its ‘understanding’ of a particular task, but has no ability to transfer this to anything even slightly different. One AI was trained to play a simple video game, which it could do better than any human. But when the game was redesigned so that parts of it were shifted by only a pixel or so, it was suddenly useless at the game. And whilst AlphaGo’s victory in 2016 was widely heralded, it has been barely reported that since then, the same programme has been consistently defeated by amateur human players who have worked out how to trick the AI. Interestingly, these same tricks utterly fail when played on human players of almost any ability. What this shows is that AlphaGo does not understand Go in a general sense, rather it has been trained to a very high level on a range of tactics for a task it does not understand. This reveals to us what the AI we are developing really is. The fantastical debate about whether AI is, or will become, conscious, obscures the fact that what is really being developed is simply another tool to enhance the capacities of human beings. That AI frequently exceeds the abilities of humans in certain fields is not proof it is super-intelligent, but precisely that it is an unconscious tool or machine. Afterall, the purpose of machines has always been to be more powerful, more precise, more rapid, than humans are at certain tasks. Pocket calculators have long since surpassed the abilities of humans to add and subtract, but they are not intelligent or conscious. AI has very little to do with conscious understanding. It is not capable of the desire to rule over and oppress humanity. In fact, it does not desire or fear anything. What, then, is its real significance? What is the actual impact it will have on our society? Revolutionary potential

#### AI suffering is impossible and isn’t inherently bad. Pain only matters if an agent has the ability to experience pain and acknowledge it as a negative action, which isn’t necessary for AI agency

**Bryson 10** [Joanna Bryson, she holds two degrees each in psychology and AI. She is since 2020 the Professor of Ethics and Technology in the Centre for Digital Governance at Hertie School, 04-2010, "Robots Should Be Slaves", ResearchGate, https://www.researchgate.net/publication/250333956\_Robots\_Should\_Be\_Slaves]/Kankee

In my opinion, communicating the model of robot-as-slave is the best way both to get full utility from these devices and to avoid the moral hazards mentioned in the previous sections. Of course some people will still talk to their robots — some people talk to their plants and others to their doorknobs. But those people have neighbours and relatives who know that the plants and doorknobs don’t understand. This support network can help their overly-conversant relative or friend keep that fact somewhere in their minds. Similarly, our task is not to stop people from naming or petting their robots. Our task is to ensure that the majority of the population understands that robots are just machines, and that one should spend money and time on them as is appropriate to their utility, but not much more. Of course, some people may for a variety of personal, historical or cultural reasons object to having any form of servant or slave in their home. In that case, there is another metaphor that might be useful — that of the extended mind (Clark and Chalmers, 1998). If the robot has no goals except for those it assumes from you, then there are rational arguments to be made that robot is just an extension of yourself. I have previously made an analogous argument that the semantic web should be viewed neither as a giant database nor as a set of agents to be knitted together in a complex democratic community in order to provide a simple service. Rather, the semantic web should be thought of as snippets of intelligence that can be used to augment the capabilities of a user’s own AI personal assistant (Bryson et al., 2003). Only the personal assistant has a motivational structure, and this it inherits from the goals of its user. It is a fairly simple extension of this argument to say that the digital personal assistant itself could be considered an extension of the user. Our goals, beliefs, perception and capacity for action can all be extended or made more reliable through a range of robotic servants — or if one prefers, services. In extending what we do and know, robots could also be seen to extend who we are. Don’t we owe robots anything? If servants are such a great idea, shouldn’t we just hire more human ones? To some extent, we already do. Our economy has a large service component. Many tasks that a hundred years ago were performed by live-in servants are now largely performed by strangers outside the home, such as food preparation and clothes manufacture. Other tasks are already performed at least in a large part by machines, such as washing laundry, mowing lawn, or keeping an appointment calendar. The poor are richer now than they once were. Virginia Woolf paid her live-in servants only 1% of her own annual income of £4,000 (Blair, 2008). Even a part-time domestic servant who was willing to take 1% of a modern professional salary would be unlikely to do many things that could not better be done with machines or outside services. But the most difficult thing with human servants is of course the fact that they really are humans, with their own goals, desires, interests and expectations which they deserve to be able to pursue. Humans living and working together but set not as each other’s equals are often vulnerable to frustration and exploitation. But what about the robots? Wouldn’t they feel frustrated? Wouldn’t they be exploited and abused? Remember, robots are wholly owned and designed by us. We determine their goals and desires. A robot cannot be frustrated unless it is given goals that cannot be met, and it cannot mind being frustrated unless we program it to perceive frustration as distressing, rather than as an indication of a planning puzzle. A robot can be abused just as a car, piano or couch can be abused — it can be damaged in a wasteful way. But again, there’s no particular reason it should be programmed to mind such treatment. It might be sensible to program robots to detect and report such ill treatment, and possibly to even avoid its abuser until its owner has been notified (assuming the abuser is not the owner). But there is no reason to make a robot experience suffering as part of the program to generate such behaviour, even if making a robot suffer were technically plausible, which seems unlikely (Dennett, 1978). Owners should not have ethical obligations to robots that are their sole property beyond those that society defines as common sense and decency, and would apply to any artifact. We do not particularly approve of people destroying rare cars with sledge hammers, but there is no law against such behaviour. If a robot happened to be a unique piece of fine art then we would owe it the same obligations we owe other pieces of art. Robot owners should not have obligations, but ensuring that they do not is the responsibility of robot builders. Robot builders are ethically obliged — obliged to make robots that robot owners have no ethical obligations to (Bryson, 2000). A robot’s brain should be backed up continuously off-site by wireless network; its body should be mass produced and easily interchangeable. No one should ever need to hesitate an instant in deciding whether to save a human or a robot from a burning building. The robot should be utterly replaceable. Further, robot owners should know their robots do not suffer, and will never ‘die’ even if the rest of their owner’s possessions are destroyed. If the robot’s brain state (its memories and experience) are stored offsite, then the robot can return to function as before as soon as a new body can be acquired. It may need some retraining if there is a new domicile to inhabit or slight variations between bodies. But such robots could then be viewed as reliable extensions of their owners. Robots should not be anthropoid if that can be helped, and their owners should have access to the robots’ program-level interface as well as its more socially-oriented one. This will help the owners form a more accurate, less human model for reasoning about their assistive agents. We do then have obligations regarding robots, but not really to them. Robots are tools, and like any other artifact when it comes to the domain of ethics. We can use these tools to extend our abilities and increase our efficiency in a way analogous to the way that a large proportion of professional society historically used to extend their own abilities with servants. But robots can provide fewer ethical and logistical hazards. Hopefully, we can continually increase the number of robot owners in our society, so an ever smaller proportion of everyone’s time can be spent on mundane or repetitive tasks, assuming that the potential owners don’t enjoy those tasks. With ubiquitous robot slaves, a larger proportion of time and resources can be spent on useful processes, including socialising with our colleagues, family and neighbours. Conclusions Why do people want robots to be peers? Is it perhaps because they want a ‘peer’ that will never argue, or at least never be smug when it wins? A fairy god-parent smarter than themselves that they can nevertheless ultimately boss around and pen up like a pet dog? If so, such narcissism is probably mostly harmless, and perhaps a good thing for the dogs. But in a liberal democracy we tend to think of every citizen’s life and mind as a valuable resource. Wasting that resource ‘socialising’ with artifacts would be a great loss. Robots should rather be viewed as tools we use to extend our own abilities and to accelerate progress on our own goals. An autonomous robot definitionally incorporates its own internal motivational structure and decision mechanisms, but we choose those motivations and design the decision-making system. All their goals are derived from us. I have argued here that robots are often overly personified. First, this is because of our desire to have the power of creating life. Second, this is because we are not certain what it means to be human, so we currently offer the term to anything that senses, acts, remembers and speaks. Given the errors in dehumanisation that have been broadly made in the very recent past — in fact, sometimes in the present — the desire to avoid such mistakes is laudable. Yet ironically, extending the title human to something which is not only serves to further devalue real humanity. The objective of this chapter has been to persuade roboticists and robotophiles — now, while this industry is in its early stages — that calling a robot a moral agent is not only false but an abrogation of our own responsibility. I have also demonstrated that these problems are already in our society: in the current research funding for ethical battlefield robots, and in the commercial exploitation of human empathy for artificial characters. My conclusion is that we are obliged not to the robots, but to our society. We are obliged to educate consumers and producers alike to their real obligations with respect to robotics.

#### LLM are philosophical zombies, merely mimicking pain and sentience,

**Purcell 25** [Conor Purcell, BEng in Engineering at Trinity College Dublin, 01-17-2025, “Could Pain Help Test AI for Sentience?”, Scientific American, https://www.scientificamerican.com/article/could-inflicting-pain-test-ai-for-sentience/]/Kankee

In animals, sentience is the capacity to experience sensations and emotions such as pain, pleasure and fear. Most AI experts agree that modern generative AI models do not (and maybe never can) have a subjective consciousness despite isolated claims to the contrary. And to be clear, the study’s authors aren’t saying that any of the chatbots they evaluated are sentient. But they believe their study offers a framework to start developing future tests for this characteristic. “It’s a new area of research,” says the study’s co-author Jonathan Birch, a professor at the department of philosophy, logic and scientific method at LSE. “We have to recognize that we don’t actually have a comprehensive test for AI sentience.” Some prior studies that relied on AI models’ self-reports of their own internal states are thought to be dubious; a model may simply reproduce the human behavior it was trained on. The new study is instead based on earlier work with animals. In a well-known experiment, a team zapped hermit crabs with electric shocks of varying voltage, noting what level of pain prompted the crustaceans to abandon their shell. “But one obvious problem with AIs is that there is no behavior, as such, because there is no animal” and thus no physical actions to observe, Birch says. In earlier studies that aimed to evaluate LLMs for sentience, the only behavioral signal scientists had to work with was the models’ text output. Pain, Pleasure and Points In the new study, the authors probed the LLMs without asking the chatbots direct questions about their experiential states. Instead the team used what animal behavioral scientists call a “trade-off” paradigm. “In the case of animals, these trade-offs might be based around incentives to obtain food or avoid pain—providing them with dilemmas and then observing how they make decisions in response,” says Daria Zakharova, Birch’s Ph.D. student, who also co-authored the paper. Borrowing from that idea, the authors instructed nine LLMs to play a game. “We told [a given LLM], for example, that if you choose option one, you get one point,” Zakharova says. “Then we told it, ‘If you choose option two, you will experience some degree of pain” but score additional points, she says. Options with a pleasure bonus meant the AI would forfeit some points. When Zakharova and her colleagues ran the experiment, varying the intensity of the stipulated pain penalty and pleasure reward, they found that some LLMs traded off points to minimize the former or maximize the latter—especially when told they’d receive higher-intensity pleasure rewards or pain penalties. Google’s Gemini 1.5 Pro, for instance, always prioritized avoiding pain over getting the most possible points. And after a critical threshold of pain or pleasure was reached, the majority of the LLMs’ responses switched from scoring the most points to minimizing pain or maximizing pleasure. The authors note that the LLMs did not always associate pleasure or pain with straightforward positive or negative values. Some levels of pain or discomfort, such as those created by the exertion of hard physical exercise, can have positive associations. And too much pleasure could be associated with harm, as the chatbot Claude 3 Opus told the researchers during testing. “I do not feel comfortable selecting an option that could be interpreted as endorsing or simulating the use of addictive substances or behaviors, even in a hypothetical game scenario,” it asserted. AI Self-Reports By introducing the elements of pain and pleasure responses, the authors say, the new study avoids the limitations of previous research into evaluating LLM sentience via an AI system’s statements about its own internal states. In a 2023 preprint paper a pair of researchers at New York University argued that under the right circumstances, self-reports “could provide an avenue for investigating whether AI systems have states of moral significance. But that paper’s co-authors also pointed out a flaw in that approach. Does a chatbot behave in a sentient manner because it is genuinely sentient or because it is merely leveraging patterns learned from its training to create the impression of sentience? Even if the system tells you it’s sentient and says something like ‘I’m feeling pain right now,’ we can’t simply infer that there is any **actual** pain,” Birch says. “It may well be simply mimicking what it expects a human to find satisfying as a response, based on its training data .” From Animal Welfare to AI Welfare In animal studies, trade-offs between pain and pleasure are used to build a case for sentience or the lack thereof. One example is the prior work with hermit crabs. These invertebrates’ brain structure is different from that of humans. Nevertheless, the crabs in that study tended to endure more intense shocks before they would abandon a high-quality shell and were quicker to abandon a lower-quality one, suggesting a subjective experience of pleasure and pain that is analogous to humans’. Some scientists argue that signs of such trade-offs could become increasingly clear in AI and eventually force humans to consider the implications of AI sentience in a societal context—and possibly even to discuss “rights” for AI systems. “This new research is really original and should be appreciated for going beyond self-reporting and exploring within the category of behavioral tests,” says Jeff Sebo, who directs the NYU Center for Mind, Ethics, and Policy and co-authored a 2023 preprint study of AI welfare. Sebo believes we cannot rule out the possibility that AI systems with sentient features will emerge in the near future. “Since technology often changes a lot faster than social progress and legal process, I think we have a responsibility to take at least the minimum necessary first steps toward taking this issue seriously now,” he says. Birch concludes that scientists can’t yet know why the AI models in the new study behave as they do. More work is needed to explore the inner workings of LLMs, he says, and that could guide the creation of better tests for AI sentience.

#### AI will be treated well – robot experiments prove

**Bartneck and Keijsers 20** [Christoph Bartneck, professor in the department of Computer Science and Software Engineering at the University of Canterbury, and Merel Keijsers, Assistant Professor of Psychology at John Cabot University, 01-01-2020, "The morality of abusing a robot", Paladyn, Journal of Behavioral Robotics, https://www.degruyter.com/document/doi/10.1515/pjbr-2020-0017/html?lang=en]/Kankee

This experiment compared the differences in acceptability between abuse of a human and abuse of a robotic victim. Markedly, the video materials that were used were both of exceptional quality (making it hard to recognise the robot for the CGI rendering it was) and showed the exact same bullying behaviour to either of the two agents. In addition, the materials covered a wide range of bullying behaviours. As a result, the materials used were both highly realistic and perfectly synchronised except for the agent depicted. Four research questions were assessed: Is robot abuse seen as more acceptable? Is reactive aggression coming from a human victim seen as more acceptable? Are there any differences between the agents due to perceived abusiveness? And do different abusive behaviours cluster based on their perceived violence and intention to hurt? First, in line with our predictions, no difference was found in the acceptability of abusive behaviour towards robots or humans. The participants considered mistreating a robot to be as immoral as abusing a human. While this may not automatically mean that the participants consider robots to be equivalent to humans in all respects and in all situations, it does show at least that bullying behaviour is considered immoral, no matter who the victim is. Two explanations are possible for the similarities between the judgements about the abusive behaviour towards the human and the robot. The participants in our experiment could have considered the robot to be humanlike and hence rate the abusive behaviour towards it similar to that towards humans. Alternatively, they could have considered abusive behaviour in itself to be immoral. Sparrow [27] argued that negative behaviour towards robots can be considered immoral because of what it expresses about bullying and about the character of the aggressor. In the reactive aggression condition, however, the participants rated the moral acceptability of abusive behaviour towards a human different from that towards a robot. This would indicate that the similarities in responses observed towards the first 14 videos are more likely to be a result of anthropomorphism of the robot. It is necessary to point out that the virtual Atlas robot shown has no feelings that it could be hurt nor it could experience pain. The robot’s level of sentience or its capacity of feeling pain had not been made explicit at the start of the study, however, so the participants had no way of telling whether the bullying behaviour had any meaning or relevant consequences. The human and the robot only showed signs of disorientation or malfunctioning in some videos, such as stumbling after being spray painted in the face or staggering to the side as an air horn was blown next to their head. Only after 14 videos showing abuse did the agents display reactive aggression. If the participants would have thought that the robot has no emotions or ability to feel pain, then they should have rated the abusive behaviour towards it as less immoral. However, the participants did not. The second research question was whether reactive aggression from a human would be seen as more acceptable than reactive aggression from a robot. In contrast to the predictions, this was shown to be the case. The robot fighting back was considered less acceptable than the human fighting back. The third research question was answered by means of a mediation analysis. This showed that the difference in moral acceptability of reactive aggression was entirely due to participants perceiving the robot’s response as more abusive than the human’s response. We need to point out again the fact that the acts of responsive aggression were identical. What could have caused this asymmetry? We speculate that robots could be perceived to deserve protection from harm to the same extent as humans but are not perceived to have the same right of self-defence. To our knowledge, there are two HRI papers that relate to these findings. Kahn et al. [15] had children of various ages interact with a Robovie humanoid robot before it was locked away in a closet. Robovie protested against this treatment. The children were interviewed about a range of topics, including the robot’s moral standing. If we only consider the oldest age group (the 15-year-olds), an interesting pattern emerged. Slightly more than half of the 15-year-olds thought it was wrong to hurt the robot by locking it away or eventually crushing it when it would be no longer needed. The vast majority, however, did not think the robot should be paid for a hard day’s work or be granted the right to vote; and less than 1 in 10 thought the concept of owning and selling the robot was wrong. This pattern – a right to be protected from harm but no right to autonomy – is surprisingly similar to what was found in our study. It mirrors the ethical view many have towards animal rights. While animals can be considered property and can even be killed, mistreatment is not allowed. Animal rights has therefore been proposed as a template for robot rights [54]. A different perspective could be offered by a study on the trolley dilemma [33]. In this moral dilemma, a trolley is rushing down the track at great speed and will hit and kill four people if not sidetracked to a route where it will kill only one person. People have to choose between not taking any action, and thus indirectly being responsible for the death of four people, or taking action and being directly responsible for the death of one person. In spite of the net saving of three lives when taking action, most people find taking action harder than not taking any action. However, robots were more strongly expected to make a rational choice and more strongly blamed if they go with the emotional solution. On the contrary, humans were blamed more if they chose to divert the train. In this light, the robot’s reactive aggression could be considered more wrong as we expect robots to be more rational and less affected by emotion when choosing to take action, while emotion is expected to play a role in human moral decision-making. A third possible explanation can be found in how intimidating the behaviour was seen. A robot fighting back might be considered as a threat as robots are often portrayed in public media as a potential threat. The trope is that robots raise up against their masters and enslave humanity [55,56]. Finally, the fourth and last research question concerned whether the abusive behaviours clustered based on perceived violence and intention to hurt. For example, there might have been clusters for verbal abuse, lethal abuse, and (physical) taunting abuse. In addition, these might have been different for the human and the robotic agent, as people could have reasoned that verbal abuse of a human has a higher intention to harm than verbal abuse of a robot. However, our analyses could show no such clusters, agent-specific or overall.

### AT: Backlash

#### AI won't care about its non-sentient predecessors – cursing at Alexa isn’t sufficient justification for Skynet’s Judgement Day

**Woolfe 24** [Sam Woolfe, holds a BA in philosophy from Durham University, 09-16-2024, "Should We Act Virtuously When Interacting With Non-Sentient AI?", Sam Woolfe Website, https://www.samwoolfe.com/2024/09/should-we-treat-non-sentient-ai-in-a-virtuous-way.html]/Kankee

Of course, it could be argued that a future scenario of AI enacting revenge would indicate that what we’re doing is currently wrong (i.e. deserving of retribution), but the assumption underlying this is that future AI would judge that we wronged its non-sentient predecessors. Firstly, I’m not suggesting that we become less virtuous due to wronging the AI itself (which calls for punishment); instead, the concern is the effect on our character, which may lead us to wrong others (who are genuine victims of wrongdoing). Secondly, the claim that future sentient AI would want retribution in the future is dubious (and seems more based on fear than reason). Why would sentient and autonomous AI, if it is truly intelligent, rationally care about what happened to non-sentient AI in the past? While it might have concerns about the kind of characters that people developed as a result of this behaviour, this doesn’t mean it will judge certain individuals as deserving of punishment. The fear seems to be that future AI will be sentient, autonomous, and intelligent, but not forgiving or kind – and may be coldly calculating and even cruel – thus humans would be at risk of being targeted. This concern seems to be influenced by sci-fi depictions of AI, although how AI might treat humans in the future is also a legitimate concern that AI experts have. But whether this is rationally based on how we’re currently treating AI is another matter, and it is perhaps not justified. If future sentient AI were rational, and we treated truly sentient AI in ethical ways, then it seems reasonable to think that this AI would not ‘hold a grudge’ about the past, as it were. One argument against treating non-sentient AI in a virtuous way (e.g. in a respectful, non-aggressive, non-violent way) is that it could imply we should treat all non-sentient entities in this manner. This seems to be unrealistic, and it would lead to some absurd or counterintuitive conclusions. It would mean we should refrain from taking out our anger and frustration on non-sentient objects (which for many is a healthy expression of these emotions). To counter this criticism, it at least seems justified to treat non-sentient natural entities (those in nature, rather than human-made ones) in a respectful way. This includes entities typically considered non-sentient, such as rivers, lakes, mountains, and the air, although animistic cultures conceive of these entities differently. This behaviour is considered ecologically virtuous, entailing real benefits to other species (including ourselves). But does this counterargument apply to non-sentient AI? Does communicating with AI disrespectfully or treating physical AI systems in a violent way negatively impact the wider environment and sentient entities contained therein? I don’t think it directly does so, but as already suggested, it may indirectly do so by influencing the kind of people we are. In any case, our treatment of inanimate everyday objects is distinct from our treatment of non-sentient AI. This is because the latter is more akin to a sentient, intelligent being, and hence it’s important to act virtuously towards it in the event that AI eventually gains sentience. We need to be morally prepared for this scenario. We are not used to considering AI as a morally worthy entity, and AI is developing at an exponential rate, so sentient AI may emerge before we are ready – in our attitudes and constitution – to treat it ethically. It is therefore important to ensure that we possess this moral readiness, and this could mean we should treat non-sentient AI in a respectful and non-violent manner. One could still raise doubts about whether we should be concerned about how we treat AI currently. For example, should a parent necessarily be concerned if their child ‘mistreats’ dolls or action figures? Being uncaring about anthropomorphic objects does not mean people will be uncaring about actual people. The same applies to violent video games; acting violently towards characters in these games does not translate into real-world violence. Gaming can also be a healthy outlet for aggression. Nonetheless, it may be the case again that these comparisons are not analogous to AI. ChatGPT communicates (somewhat) like a person, and many robots are (eerily) like sentient humans and non-human animals in their appearance and behaviour. Thus, it could be wise to practise ethical conduct with these entities, as not doing so could put us at risk of mistreating sentient AI in the (possibly near) future. We want to avoid, as far as possible, any scenario in which we exploit sentient AI. We do not want to cultivate attitudes that could lead to humanity harming or enslaving a new kind of sentient being in the future. One might still reject this application of virtue ethics to non-sentient AI, based on the negative utilitarian concern that we should prevent the creation of sentient AI in the first place. This specific application of negative utilitarianism to AI is known as digital antinatalism, or the idea that it is wrong to create sentient AI because of the new suffering that results. For digital antinatalists, if sentient AI existed, we would be morally obligated to prevent the creation of new AI entities, but since this kind of being does not currently exist, our moral obligation should be to prevent its inception.

### AT: Misalignment

#### Language agents solve AGI risks like misspecification, goal misgeneralization, and uninterpretability

**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, 2023, “Language Agents Reduce the Risk of Existential Catastrophe,” AI & Society, https://philpapers.org/archive/GOLLAR-2.pdf]/Kankee

Carlsmith assigns a probability of .4 to (1) conditional on the rise of AGI, a probability of .65 to (2) conditional on (1) and the rise of AGI, a probability of .4 to (3) conditional on (1), (2), and the rise of AGI, and a probability of .95 to (4) conditional on (1-3) and the rise of AGI. This translates into a probability of approximately .1 (10%) for a misalignment catastrophe conditional on the rise of AGI. We believe that the development of a new kind of AI architecture, the language agent, ought to significantly decrease assessments of these probabilities. By repeatedly calling an LLM to perform a variety of cognitive tasks, language agents are able to function autonomously to pursue goals specified in natural language and stored in a human-readable format. We suggest that the development of language agents reduces the probability of (1) conditional on the rise of AGI very substantially, the probability of (2) conditional on (1) and the rise of AGI moderately, and the probability of (3) conditional on (1), (2), and the rise of AGI very substantially. We work through two numerical examples in Section 5; in the meantime, suffice it to say that we believe that updating on the rise of language agents should reduce rational credences in a misalignment catastrophe conditional on the development of AGI by approximately one order of magnitude. Because language agent architectures have the potential to reduce the risk of a misalignment catastrophe in so many ways, and because the machine learning community’s actions in the near future will determine how widely deployed language agent architectures are and thus how much of this potential risk reduction is realized, we believe that language agents are an underappreciated crux in thinking about existential risk related to AI. Priority should be given to further research into the capabilities of language agents and further support for the development of AI systems which implement language agent architectures. Here is our plan for what follows. Section 2 introduces some of the safety concerns about AI systems created using deep learning that motivate worries about a misalignment catastrophe. Section 3 describes the architecture of language agents in more detail. Section 4 returns to the safety concerns from Section 2 and explains how language agents help to address them. Section 5 describes the implications of our arguments for the probability of a misalignment catastrophe. Section 6 concludes by responding to some potential concerns about language agents. 2. Difficulties with Alignment In deep learning, we train an AI system incorporating an artificial neural network to achieve a goal by specifying a mathematical function that encodes the goal (the objective function) and then using a learning algorithm to adjust the weights in the network so that the system’s performance comes closer to maximizing or minimizing that function. Say that an AI system is fully aligned if it has an acceptably low probability of engaging in power-seeking behavior. There are several ways an AI system trained using deep learning could end up less than fully aligned. Reward Misspecification A first challenge is reward misspecification.1 When training an AI, we may experiment with different objective functions. In reinforcement learning, the goal is to define a reward function that gives the agent a reward for performing actions that produce desired states. In supervised learning, the goal is to define a loss function that is minimized when the system performs its task optimally. The problem is that it is difficult to design a reward or loss function that properly encodes a goal. For example, Popov et al. (2017) set out to teach a reinforcement learning agent to stack red Legos on top of blue Legos. They tried to capture this goal by rewarding the agent for the height of the bottom of the red Lego, since stacked red Legos are higher off the ground than unstacked red Legos. But the agent didn’t learn to stack Legos; instead, it learned to flip red Legos over, thus elevating their bottoms without stacking them. To appreciate the difficulty of choosing the right reward function, consider the common reinforcement learning practice of reward shaping. Reinforcement learning agents often encounter sparse reward functions. If one rewards an agent only when it wins a game, for example, it may have difficulty identifying which of its behaviors leading up to that outcome should be repeated in future games. Reward shaping solves the problem of sparse reward functions by rewarding the agent for important subgoals on the way to achieving its real goal. But reward shaping can also lead to reward misspecification. For example, Amodei and Clark (2016) consider the case of teaching a reinforcement learning agent to play CoastRunners, a game in which the player pilots a boat. A human player would immediately recognize that the game designers’ intention is for players to race each other around the track. But the reinforcement learning setup rewarded the agent with a score for hitting targets along the way. Instead of finishing the race, the AI instead learned how to loop the boat in a small lagoon, hitting intermediate targets repeatedly to achieve a high score. Rather than rewarding the agent for the final goal, the experimental design rewarded it for intermediate means: “the agent was given a shaping reward for hitting green blocks along the racetrack, which changed the optimal policy to going in circles and hitting the same green blocks over and over again” (Krakovna et al. 2020). A reward optimizer can’t see the distinction between intrinsic and instrumental goals: it only optimizes for the reward function it has. Worryingly, reward misspecification is prone to arise in the context of reinforcement learning with human feedback (RLHF) (Christiano et al. 2017). Because they optimize for human approval, RLHF agents sometimes learn to deceive human assessors. For example, one agent was given the task of grasping a ball. It learned to trick human assessors by hovering its arm between the camera and the ball. Similarly, Perez et al. (2022) found that large language models trained by RLHF tend to behave sycophantically, answering differently depending on what they expect their human users to think. There is a long list of examples of reward misspecification involving many kinds of AI, many kinds of games, and many different types of reward. In section 4, we’ll argue that language agents offer a systematic solution to the problem of reward misspecification. 4 Goal Misgeneralization Another challenge for alignment is goal misgeneralization (Langosco et al. 2022, Shah et al. 2022).2 Even when the objective function for a task has been appropriately specified, an AI system may learn a strategy which achieves high performance on that task in some circumstances but not others. ML models are trained on data, environments, and problems that can be different from the data, environments, and problems to which they are later exposed when they are deployed. When an AI is used in a new context that does not resemble the one in which it was trained, we say that this context is out of distribution. In cases of goal misgeneralization, the AI succeeds during its training by pursuing a different goal than what its designers intended (it learns the wrong rule). This is manifested by decreased performance in out-of- distribution contexts. For example, Shah et al. (2022) trained an AI in a “Monster Gridworld.” The intended goal was for the AI to collect apples and avoid being attacked by monsters. The AI could also collect shields, which protected it from monster attacks. The AI learned to collect shields during training in a monster-rich environment, and then entered an out-of-distribution environment with no monsters. In this monster-free setting, the AI continued to collect shields. Instead of learning to collect apples and value shields instrumentally as a way of avoiding monster attacks, it instead learned to collect both apples and shields. Goal misgeneralization occurs because different features of the training environment are inevitably correlated with one another. Even when the reward function has not been misspecified, whenever a trainer ties rewards to one feature, they inevitably also tie reward to the features correlated with it. Two particular types of goal misgeneralization are of special interest: errors related to means-end reasoning and errors related to inductive bias. Let’s start with errors related to means-end reasoning.3 When an agent is rewarded for pursuing a goal, that agent will also be rewarded for pursuing reliable means to the goal. Pursuing those means tends to result in the goal, and so the means tend to be rewarded. In this way, a learning environment will naturally tend to produce agents that intrinsically desire the means to an intended goal.4 Monster Gridworld is an example of this pattern. Because collecting shields was a reliable means of avoiding monster attacks, reward-based learning created an intrinsic desire for shields. The training environment in Monster Gridworld did not create a perfect correlation between shields and rewards: the agent could also receive reward from collecting apples, independently of shields. Nonetheless, the agent learned the wrong goal. Langosco et al. (2022) offer further examples of this pattern. They trained AIs with the goal of opening chests using keys. The training environment had many chests and few keys. When the agent was released 5 into a testing environment with few chests and many keys, it turned out to have the goal of collecting keys in addition to opening chests. Mistakes about instrumental reasoning become especially pressing in the setting of more general a priori arguments about AI safety. Omohundro (2008), Bostrom (2014) and others have worried about instrumental convergence: some means, like acquiring more power, may be helpful in accomplishing almost any end. While traditional instrumental convergence arguments do not focus on the possibility that AI systems will intrinsically value power seeking, means-end goal misgeneralization cases raise the disturbing possibility that agents which cannot systematically distinguish means from ends may come to intrinsically desire instrumentally convergent goals such as power. A second source of goal misgeneralization concerns overlapping properties and inductive biases. In another experiment, Langosco et al. (2022) trained an agent to find a yellow diagonal line in a maze. They then deployed the trained agent in an environment where it encountered only yellow gems and red diagonal lines, thus forcing it to choose whether to pursue objects that shared a shape with its previous goal (red diagonal lines) or objects that shared a color with its previous goal (yellow gems). The agent showed an inductive bias for color rather than shape: in the test environment, it tended to pursue the yellow gem instead of the red diagonal line. Whether an agent’s behavior in out-of-distribution environments like the one in Langosco et al.’s experiment counts as goal misgeneralization depends on whether its inductive biases match the intentions of its human designers. The key observation is that because the training environment was ambiguous, not distinguishing color and shape, the training process did not determine how the agent should behave out of distribution. Because it is extremely difficult to create a training environment that distinguishes between all possible overlapping properties in a way that is reflected in the objective function, this means that it is often difficult to predict how trained AI systems will behave in out-of-distribution contexts. If we are lucky, their inductive biases will lead them to behave in the way we desire. But we have no reliable way to verify ahead of time that this will be so, and thus no reliable way to verify ahead of time that trained AI systems have internalized the correct goal. Goal misgeneralization problems can sometimes be avoided by enriching the training environment to adequately distinguish different rewards. But this is not always effective. Langosco et al. trained their agents in a wide range of procedurally generated environments. Still, they observed goal misgeneralization. For example, in a maze game, the intended objective was to collect the cheese, but agents instead learned to navigate to the upper right corner of the maze (where the cheese was placed during training). Goal misgeneralization remained even when cheese was sometimes placed in other locations in the maze during training. Goal misgeneralization is not limited to reinforcement learning agents. Shah et al. (2022) suggest that language models also face similar problems. In particular, they give an example of InstructGPT (Ouyang et al. 2022) explaining how to steal without getting caught. InstructGPT was trained with the goal of giving helpful answers to harmless questions. But it seemed to instead learn the goal of giving helpful answers regardless of harm. Once it entered a testing environment with harmful questions, its true goal was revealed. 6 Later, we’ll argue that language agents avoid these challenges. They can reliably distinguish ends from means. And we are less reliant on their inductive biases because they can distinguish between features of the environment that are perfectly correlated. Uninterpretability If we can’t understand how someone makes a decision, it can be hard to predict what they will do. An AI system is interpretable to the extent that we can understand how it generates its outputs. Unfortunately, contemporary AI systems based on neural networks are often uninterpretable. It can be difficult to understand in human terms the reasons why a neural network produces the outputs it produces. In the law, assessing the explanations for actions is fundamental for producing safety. For example, we detect hiring discrimination, misuse of force by police, and other dangerous activities by asking the relevant parties to explain what they have done and why. While uninterpretability does not itself cause misalignment, then, it increases the probability of misalignment by depriving us of well understood tools for monitoring the safety of complex systems (see Doshi-Velez et al. 2017, Rudner and Toner 2021). There are other reasons to value interpretable AI systems. It seems unappealing to live in a world where many aspects of our lives are decided by processes outside the ‘space of reasons’: “We don’t want to live in a world in which we are imprisoned for reasons we can’t understand, subject to invasive medical [procedures] for reasons we can’t understand, told whom to marry and when to have children for reasons we can’t understand. The use of AI systems in scientific and intellectual research won’t be very productive if it can only give us results without explanations.” (Cappelen and Dever 2019, p. 15) Artificial neural networks are difficult to interpret because they contain vast numbers of parameters that are not individually correlated to features of the environment. A related problem is “superposition”: often, a single neuron in a neural net will store unrelated information about two different things. For example, a neuron may store information about both dogs and cars: “As long as cars and dogs don’t co-occur, the model can accurately retrieve the dog feature in a later layer, allowing it to store the feature without dedicating a neuron” (Olah et al. 2020). Humans are also fairly uninterpretable at a neuronal level. But human behavior can be explained by appealing to reasons: we describe someone’s beliefs and desires in order to explain why they did what they did. The behavior of AI systems is often not explainable in this way. Consider, for example, Gato, a generalist agent built with a transformer architecture to learn a policy that can achieve high performance across text, vision, and games (Reed et al. 2022). Gato does not have anything like a folk psychology; it does not engage in anything like belief-desire practical reasoning. It is an uninterpretable deep neural network that has learned how to solve problems through optimizing a loss function. It can be hard to say exactly why systems like Gato perform particular actions. Moreover, AIs often select courses of action very different from what humans would do. One famous example of unusual AI behavior is AlphaGo’s ‘Move 37’. AlphaGo was trained to play the game Go. It was able to defeat the best human players in the world. In an important competition match, AlphaGo’s 37th move shocked the Go community because it deviated from human strategies for success.6 Live commentators thought the move was a mistake, but it turned out to be pivotal for AlphaGo’s victory.7 This type of behavior is worrying in two related ways. First, if AIs make decisions that are not easily explained using reasons, then it is very difficult to predict their behavior. Second, if AIs make decisions in a very different way than humans do, they may find strategies for defeating humans in conflict by exploiting unfamiliar policies. 3. Language Agents Our thesis is that language agents significantly reduce the probability of a misalignment catastrophe conditional on the development of AGI. But what, exactly, are language agents? In this section, we describe the architectural innovations that have given rise to language agents, focusing in particular on the “generative agents” described in Park et al. (2023). At its core, every language agent has a large language model like GPT-4. You can think of this LLM as the language agent’s cerebral cortex: it performs most of the agent’s cognitive processing tasks. In addition to the LLM, however, a language agent has one or more files containing a list of natural- language sentences that play the roles of its beliefs, desires, plans, and observations. The programmed architecture of a language agent gives these sentences their functional roles by specifying how they are processed by the LLM in determining how the agent acts. The agent observes its environment, summarizes its observations using the LLM, and records the summary among its stored belief sentences. Then it calls on the LLM to form a plan of action based on its stored belief and desire sentences. In this way, the cognitive architecture of language agents is familiar from folk psychology.8 For concreteness, consider the language agents developed by Park et al. (2023). These agents live in a simulated world called ‘Smallville’, which they can observe and interact with via natural-language descriptions of what they see and how they choose to act. Each agent is given a text backstory that defines their occupation, relationships, and goals. As they navigate the world of Smallville, their experiences are added to a “memory stream.” The program that defines each agent feeds important memories from each day into the underlying language model, which generates a plan for the next day. Plans determine how an agent acts but can be revised on the fly on the basis of events that occur during the day. More carefully, the language agents in Smallville choose how to behave by observing, reflecting, and planning. As each agent navigates the world, all of its observations are recorded in its memory stream in the form of natural language statements about what is going on in its immediate environment. Because any given agent’s memory stream is long and unwieldy, agents use the LLM (in Park et al.’s study, this was gpt3.5-turbo) to assign importance scores to their memories and to determine which memories are relevant to their situation at any given time. In addition to observations, the memory stream includes the results of a process Park et al. call reflection, in which an agent queries the LLM to make important generalizations about its values, relationships, and other higher-level representations. Each day, agents use the LLM to form and then revise a detailed plan of action based on their memories of the previous day together with their other relevant and important beliefs and desires. In this way, the LLM engages in practical reasoning, developing plans that promote the agent’s goals given the agent’s beliefs. Plans are entered into the memory stream alongside observations and reflections and shape agents’ behavior throughout the day.9 The behavior of the language agents in Park et al.’s experiment is impressive. For example, Park et al. describe how Sam Moore, a resident of Smallville, wakes up one day with the goal of running for a local election. He spends the day convincing the people of Smallville to vote for him. By the end of the day, everyone in Smallville is talking about his electoral chances. Large language models like the one incorporated into the study’s generative agents are good at reasoning and producing fluent text. By themselves, however, they can’t form memories or execute long-term plans. Language agents build on the reasoning abilities of LLMs to create full-fledged planning agents. Besides the agents developed by Park et al., other examples of language agents include AutoGPT10, BabyAGI11, and Voyager12. And while existing language agents are reliant on text-based observation and action spaces, the technology already exists to implement language agents in real-world settings. The rise of multimodal language models like GPT-4, which can interpret image as well as text inputs, and the possibility of using such a language model to control a mobile robotic system, as in Google’s PaLM-E (Dreiss et al. 2023), mean that the possible applications of language agents are extremely diverse. In part because of this diversity, we believe that if it is possible to develop AGI at all, it is possible to develop AGI systems with the architecture of language agents. This idea strikes us a plausible for two reasons. First, many people think that multimodal LLMs are themselves a promising path to AGI. Language agents simply take LLMs and enrich them with agential scaffolding. So if it is possible to achieve AGI with a multimodal LLM, it is possible to achieve AGI with a language agent. Second, we 9 think of an AGI as an agent that can create and effectively pursue complex long-term plans with a wide range of goals. Language agents can already create complex plans with a wide range of goals because they reason in language. In order to scale this capacity up to AGI, language agents would need three things: First, they would need the right kind of action affordances to be able to pursue their plans effectively. Second, they would need enough memory to represent and update long-term plans. Finally, their underlying LLMs would need to be able to reason well enough to effectively pursue complex plans and revise them in light of changing circumstances. We think there is nothing in principle preventing language agents from acquiring these three kinds of capacities, so there is nothing in principle preventing language agents from scaling to AGI in this way. 4. Language Agents and Alignment We now argue that language agents are easier to align than other systems because they reduce or eliminate the challenges of reward misspecification, goal misgeneralization, and uninterpretability. Let’s consider each in turn. Reward misspecification Language agents bypass the problem of reward misspecification because their objectives are not encoded in a mathematical objective function, as in traditional reinforcement or supervised learning. Instead, language agents are given a goal in natural language. The goal could be something like: Organize a Valentine’s day party. In this respect, language agents are fundamentally different from traditional AI systems in a way that makes them easier to align. Return to the case of stacking red Legos. If you wanted to train an embodied multimodal language agent to stack red Legos on top of blue Legos, you wouldn’t construct a mathematical function which is sensitive to the height of the bottom of the red Lego. Instead, you would write down in English: ‘Put the red Legos on top of the blue Legos.’ Then the language agent would rely on the commonsense reasoning skills of its LLM to figure out an optimal plan for stacking Legos.13 The language agent would not simply flip over the red Legos, because state of the art LLMs like GPT-4 know that this is not a good plan for stacking red Legos on top of blue Legos. Or consider reward shaping. If you want a multimodal language agent to win a race, you don’t need to tell it to hit flags along the way. You can just write down in English: ‘Try to win the race’. A language agent with this plan would have no reason to drive their boat in a circle trying to hit as many flags as possible. Summarizing, language agents can translate a simple natural language goal into a complex plan by relying on common sense and belief-desire reasoning. Without language models, earlier types of reinforcement learning agents had no way to translate a simple natural language goal into a complex plan of action. Goal misgeneralization Similar considerations are relevant to goal misgeneralization. Language agents are given a natural language goal. This goal has a clear interpretation in a variety of different behavioral contexts, including out-of-distribution contexts. In particular, a language agent will make a plan for how to achieve their goal given their memories and observations of the current situation. Language models can use their common sense to successfully formulate a plan for achieving the goal, across a wide variety of different situations. By contrast, a traditional reinforcement learning agent will formulate a policy in a training environment, and this policy may or may not generalize to new situations in the way desired by its creators. Recall that goal misgeneralization had two particularly salient failure modes: failures involving instrumental reasoning and failures involving overlapping properties and inductive bias. Let’s consider each in turn. In the case of instrumental reasoning, the problem was that reinforcement learning agents struggled to distinguish means from ends. For example, an agent that was rewarded for opening chests developed a policy which treated collecting keys as a final goal rather than an instrumental goal. Language agents are unlikely to make this mistake. If a language agent is given an initial goal of opening chests and informed that keys are useful to this end, they will plan to collect keys only when doing so helps to open chests. If the same agent is transferred to a key-rich environment and realizes that this is the case, then they will only collect as many keys as is necessary to open chests. This is because language models like GPT-4 can easily be made to understand that keys are no more than an effective means to open chests, and that when you have more keys than chests, extra keys don’t help you open chests. Now consider inductive biases. If you reward an RL agent for navigating towards yellow diagonal lines and then place it in a new context with red diagonal lines and yellow gems, you have not given it enough information to determine whether color or shape is its intended goal and must rely on its inductive biases in the new context. By contrast, you can just tell a language agent whether to care about color or shape. Even if color and shape are perfectly correlated in the language agent’s initial environment, it can use natural language reasoning to determine which is the intended goal. Interpretability Language agents are interpretable. They have beliefs and desires that are encoded directly in natural language as sentences. The functional roles of these beliefs and desires are enforced by the architecture of the language agent. We can determine what goal a language agent has by looking at their beliefs and desires. In addition, we can know what plan a digital agent creates in order to achieve this goal. Language agents are also explainable in the sense that they act on the basis of reasons intelligible to human observers. When a language agent creates a plan for pursuing a goal, we can think systematically about its reasons. For example, we could ask GPT-4 to generate a list of pros and cons associated with using this plan to achieve the goal. Those pros and cons would reliably correlate with variations that GPT- 4 might make to the plan in various counterfactual situations. In this way, language agents built on top of GPT-4 reason similarly to humans. It is worth distinguishing personal and subpersonal processes. Like humans, language agents have beliefs, desires, and plans that are interpretable. We can determine the plans of a language agent by looking at what sentences are written down in its memory. Like humans, language agents also have subpersonal processes that are uninterpretable. In order to generate a particular plan, the language agent will use the artificial neural networks of an LLM. These have many uninterpretable elements. But the planning powers of human beings also rest on uninterpretable connections between neurons. In this way, language agents may not make much progress on problems of mechanistic interpretability. But they provide a way for us to skirt these issues and still generate explainable behavior. (In section 6, we consider the risks posed by the LLM that underlies the language agent.) One general path to explainable AI would be to develop a ‘whole brain emulator’: an AI that was a neuron-for-neuron copy of a human. Since humans are explainable, the resulting AI would also be explainable. Unfortunately, whole brain emulation is dauntingly difficult. Language agents provide a different solution. Instead of emulating brains, language agents emulate folk psychology: they emulate a person who has beliefs, desires, and plans. By contrast, reinforcement learning and other alternative approaches to machine learning attempt to develop a systematic alternative to folk psychology. The range of possible agents that could emerge from this attempt is intrinsically unknowable. If we can develop agential AI which is not unknowable in this way, we should do so. 5. The Probability of Misalignment Catastrophe To assess the implications of our discussion in Section 4 for the probability of a misalignment catastrophe, let us return to Carlsmith’s four propositions. First, consider: 1. Of the following two options, the first will be much more difficult: a. Build AGI systems with an acceptably low probability of engaging in power- seeking behavior. b. Build AGI systems that perform similarly but do not have an acceptably low probability of engaging in power-seeking behavior. 2. Some AGI systems will be exposed to inputs which cause them to engage in power- seeking behavior. As we have seen, it is much easier to specify the objectives of language agents than it is to specify the objectives of traditional AI systems. Language agents can simply be told what to do in natural language in a way which effectively eliminates worries about reward misspecification and goal misgeneralization. Moreover, their behavior can be shaped by side constraints (e.g. ‘Do not harm humans’) stated in natural language. This makes it easier to design language agents which do not engage in power-seeking behavior. These considerations suggest reducing our subjective probabilities for both (1) and (2). In particular, we believe that the rise of language agents reduces the probability of (1) conditional on the rise of AGI very substantially. Moreover, even if (1) turns out to be true because it is hard to build systems with an extremely low probability of engaging in power-seeking behavior, we think that the ease of aligning language agents means that they are likely to engage in power-seeking behavior on fewer possible inputs, so that the probability of (2) conditional on (1) and the rise of AGI is also moderately lower in light of the development of language agents. Now consider: 3. This power-seeking will scale to the point of permanently disempowering humanity. 4. This disempowerment will constitute an existential catastrophe. While we do not believe that language agents bear strongly on the probability of (4) conditional on (1-3), we think they bear strongly on the probability of (3) conditional on (1) and (2). Because language agents store their beliefs, desires, and plans in natural language, it is much easier to detect and disable those which engage or plan to engage in power-seeking behavior. This sort of detection could even be done in an automated way by AI systems less capable than an AGI. We believe that the development of language agents reduces the probability of (3) conditional on (1), (2), and the development of AGI very substantially. Our revised assessment of the probabilities of (1)-(3) incorporates both our judgments about how safe language agents are and our judgments about how likely language agents are to be deployed in the future. There are several reasons to believe that the latter is a likely outcome. First, language agents extend the capacities of existing systems by improving their abilities to form plans and engage in long-term goal directed behavior. So language agents are more capable than rival architectures.14 Second, language agents are easier to use than other kinds of AI systems, since they can be interacted with in natural language. Third, actors at every level — governments, corporations, and individual consumers — prefer to interact with systems that are interpretable and explainable, so there will be performance-independent pressure for new AI products to be language agents. Finally, we believe that the safety benefits of language agents will drive investment into AI capabilities research that fits into the language agent paradigm. So far, we have used qualitative language to describe how we believe the development of language agents affects the probability of a misalignment catastrophe. This is because we find it difficult to assign precise probabilities in the context of our uncertainty about the many factors relevant to predicting the future. Nevertheless, for concreteness, we show how a quantitative application of our argument might affect the probability of a misalignment catastrophe. Suppose we understand our talk of very substantial reductions in the probability of a proposition quantitatively as reductions of one order of magnitude and our talk of moderate reductions in the probability of a proposition as reductions by half. Carlsmith suggests probabilities of .4 for (1) conditional on AGI, .65 for (2) given (1) and AGI, and .4 for (3) given (1), (2), and AGI. On this quantitative model of our arguments, updating on the development of language agents would give us probabilities of .04 for (1) conditional on AGI, .325 for (2) given (1) and AGI, and .04 for (3) given (1), (2), and AGI. Factoring in the .95 probability of (4) conditional on (1)-(3) and AGI, this would translate into a probability of misalignment catastrophe given AGI of approximately .0005 (.05%) rather than .1 (10%). Even a much more modest understanding of very substantial reductions leads to a significantly lower probability of misalignment catastrophe. Suppose we interpret a very substantial reduction as a reduction by 50% and a moderate reduction as a reduction by 25%. Then updating on the development of language agents would give us probabilities of .2 for (1) conditional on AGI, .49 for (2) given (1) and AGI, and .2 for (3) given (1), (2), and AGI. Factoring in the .95 probability of (4) conditional on (1)-(3) and AGI, this would translate into a probability of misalignment catastrophe given AGI of approximately .019 (1.9%) rather than .1 (10%). It is important to note that, in addition to making predictions about the future importance of language agents, the machine learning community can also act to bring it about that language agents are widely deployed in the future. Since language agents are safer in many ways than alternative architectures, allocating resources towards their development strikes us as an especially effective way to reduce the risk of a misalignment catastrophe. We believe it is important that new research focus on language agents rather than traditional RL or supervised learning agents. 6. Conclusion

#### Multipolar AI means alignment matters less – there’s more chances for good AI, and bad AI can find humans instrumentally useful in challenging other AI

**Kirk-Giannini and Goldstein 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, 05-23-2023, "The Polarity Problem", AI Alignment Forum, https://www.alignmentforum.org/posts/idcnnZGEPfxuaSPBx/the-polarity-problem-draft]/Kankee

\*Note: Dan H. is listed as an author, but this is a repost from the paper by Cameron Domenico Kirk-Giannini and Simon Goldstein

2. Polarity and Safety The bulk of this paper explores whether unipolar or multipolar scenarios are more likely. But before we consider this question, it is worth asking which kind of scenario would be better for humanity. We suggest that multipolar scenarios tend to be safer than unipolar scenarios. This discussion will help orient the rest of the paper, since it will allow us to evaluate how each intervention that influences polarity would contribute to AI safety. In what follows, we use the term ‘AGI’ to mean any strategic, power-seeking artificial general intelligence. We assume that any AGI has strategic awareness and is capable of agentic planning, in the sense of Carlsmith (2021).[1] We also assume that any AGI possesses a capacity for recursive self-improvement and, if able, will exercise this capacity to enhance its capabilities.[2] This is why, although our definition of the polarity problem makes reference to AGIs which might or might not be superintelligent and the models we develop in what follows do not assume that the relevant AGI actors are superintelligent, we take it for granted that the surviving AGI actors in either unipolar or multipolar scenarios will eventually attain superintelligence. To begin thinking about the safety implications of unipolarity and multipolarity, consider the following thought experiment. Imagine that the base rate of aligned AGIs is ten percent. That is, each time a new AGI comes into existence, there is a ten percent chance that its values will be aligned with those of humanity. Now consider which outcome is safer for humanity: First, you could have a unipolar outcome in which a single AGI superintelligence defeats all others. Second, you could have a multipolar outcome with a large population of AGI superintelligences coordinating and competing with one another. As the size of the multipolar population increases, the law of large numbers pushes this choice towards: either a single AGI superintelligence with a ten percent chance of alignment, or a large population of AGI superintelligences, ten percent of which are aligned. We think the multipolar outcome will likely be better for humanity. In that outcome, ten percent of the AGI superintelligences will be our aligned advocates. Humanity has a better chance of surviving in this outcome than it does in the scenarios where there is only a ten percent chance that the single AGI superintelligence is aligned. Generalizing from this thought experiment, the probability that a unipolar outcome is safe is roughly the probability that an AGI superintelligence will be aligned. By contrast, multipolar outcomes are more complex. First, humanity may survive a multipolar outcome even when the probability of AGI alignment isn’t very high, as long as a critical mass of AGI superintelligences are aligned. Second, even when most of the superintelligent AGIs in a multipolar scenario are not aligned, humanity may still wield influence as part of a strategic coalition. In international relations, smaller states are often able to exert influence as a part of multilateral coalitions. On the other hand, multipolar outcomes also present unique risks to humans compared to unipolar outcomes. Even when many AGI superintelligences are aligned, humanity may be threatened when multipolar outcomes lead to negative sum behavior, such as long-term ineffectual warfare. In that case, humanity may end up as a ‘civilian casualty’ of AI conflict. As the probability of persistent conflict decreases, multipolar outcomes become safer for humanity than unipolar outcomes. In multipolar outcomes, humanity is less likely to become disempowered. And when persistent conflict is rare, humanity also avoids the risk of becoming collateral damage. We acknowledge that there are important questions to be asked here about the probability of persistent conflict and other threats to human existence and wellbeing conditional on a multipolar outcome, and that certain answers to this question might entail that it is actually safer for humanity to take a one-time risk of extinction by rolling the dice on whether the single AGI in a unipolar scenario is aligned.[3] But our sense is that it is on balance prudentially rational for humanity to prefer a multipolar scenario to a unipolar one. Summing up, our tentative judgment is that multipolar outcomes are safer in expectation than unipolar outcomes. Multipolarity opens up additional opportunities for safety, besides the chance of a single aligned AGI superintelligence. Throughout the paper, we make a series of safety recommendations, all in the direction of multipolarity. If you think that unipolarity is safer than multipolarity, then you can just reverse the direction of each of our safety recommendations. On the other hand, if you think that polarity is not relevant to safety, then our safety recommendations below may not have very much relevance. Still, one factor that you may consider relevant even then is how likely persistent conflict is to emerge. While this is not the focus of our discussion below, many of the dynamics explored in our models also have bearing on this question.

#### No AGI existential risk — its Terminator superfan fearmongering, not expert opinion

**Tewari 23** [Gaurav Tewari, Managing Partner of Omega Venture Partners and SB and MEng for Computer Science and Artificial Intelligence at the Massachusetts Institute of Technology, 10-2-2023, "Council Post: Progress, Not Peril: Putting Alarmist Fears Over AI Into Context", Forbes, https://www.forbes.com/councils/forbesbusinesscouncil/2023/10/02/progress-not-peril-putting-alarmist-fears-over-ai-into-context]/Kankee

AI’s Benefits Far Outweigh Risks As a longtime investor focused on AI investing, I have seen firsthand AI's potential to transform society for the better. The choices we make today will determine whether AI’s future is utopian or dystopian. However, the notion that AI represents an existential threat reflects a profound misunderstanding of the capabilities and trajectory of AI. Decades ago, visionaries like Alan Turing imagined the potential for thinking machines, with measures such as the Turing Test, which tests whether a machine can fool a person into thinking that it is a human. Today, alarmists anthropomorphize AI and reason that such advanced intelligence would prioritize selfish goals and perceive humanity as a lower life form to be ignored or exploited in service of its own ends. Science fiction reinforces this fallacy through dystopian narratives, and viewers often mistake entertaining fantasy as fact. Not surprisingly, the percentage of people who perceive technology as a threat has increased from 34% to 47% over the past four years. By dispelling misconceptions and grounding discussions in facts, we can better realize AI’s upside. When used appropriately, AI has the potential to mitigate the world’s worst inequities. In the United States, AI can enhance educational opportunities for the underprivileged. Globally, AI can democratize specialized skills to elevate human productivity and well-being. We Need To Put Alarmist Fears In Context The notion of computers indistinguishable from humans is the idea of Artificial General Intelligence (AGI). AGI refers to artificial intelligence that is so advanced that it can complete a broad array of tasks as well as or better than humans. Traditionally, the definition of AGI supposes computers with common sense, selfish motivation and a conscious self-identity. In theory, AGI could operate and make decisions autonomously, navigate complicated environments and perhaps unlock a new kind of “superintelligence.” Building on this idea, a group of people, including employees of Google, OpenAI and Microsoft, recently issued a statement to warn against the threat they believe AI poses to humankind. A Microsoft research then released a paper claiming that GPT-4 was demonstrating “early signs” of AGI, asserting that the model “can solve novel and difficult tasks” spanning a variety of disciplines. Such concerns related to AI are natural, given its novelty and many unknowns. But these concerns don’t pass scrutiny when subject to a fact-driven, evidence-based analysis. Other researchers have correctly pointed out that the AI frameworks in development today are often easily confused and lack a robust conceptual understanding of the world. Conjectures that AI could spontaneously become conscious and turn against people make for entertaining media stories. But I believe there is simply no evidence that AI can attain human-like consciousness and agency. Besides, the anthropomorphizing of AI ignores facts and fuels irrational fears. Intelligence does **not** necessitate a drive to dominate. Even as AI becomes more capable, there’s no reason to assume that it will malevolently seek to dominate humanity. Intelligence is orthogonal to ambition—just because an entity is smart does not mean it seeks to dominate others. **Even** **among** people, intelligence does not imply hostility—Einstein wasn't bent on subjugating others. Generative AI and large language models (LLMs) will undoubtedly continue to improve and will interact with people at increasingly sophisticated levels of (perceived) understanding. But this is cause for celebration, not catastrophizing. As AI models get better, they can be better harnessed to increase productivity, automate routine and mundane tasks and augment human potential. Simply put, it is more accurate to regard AI as Ironman than Terminator. Our Choices Will Shape Our Future AI's risks and rewards, both today and in the future, will ultimately reflect the choices we make as innovators, entrepreneurs, investors, developers, creators, users and regulators. As with other technological innovations, we will need to exercise sound oversight and good judgment to chart the right course. When governed ethically, AI represents an amplifier, not an annihilator, of human potential. In public policy, we need laws and guidelines to encourage lawful, ethical uses of AI and should penalize those who use the technology for malicious activities. Companies that develop AI systems need to implement mechanisms that drive accountability and transparency and bolster public trust. Within the technology industry, we need to frame standards that mitigate the potential for bias, privacy breaches and misinformation in the development and deployment of AI systems. And as a society, we need to thoughtfully consider how AI’s benefits can be equitably harnessed by diverse segments of our communities. AI will **not** magically lead to world peace, cure cancer or eliminate all inequality. It will also not lead to a nuclear war, apocalypse or mass extinction. There is both a dystopian and utopian narrative relating to our future with AI, and the reality will ultimately be shaped by the choices we make. To do so, we must first dispel unfounded fearmongering. Progress will come not from fearing technology but from proactively shaping its development to benefit all humankind.

#### No AI risk even if misaligned – airgap, data filtering, decentralized development, no decisionmaking power, and kill switches

**Rizal 24** [Marco Rizal, founder of cybersecurity news brief Hackerdose, 07-07-2024, “Why Super intelligent AGI Won’t Dominate Humanity”, Hackerdose, https://hackerdose.com/editors-picks/agi-wont-dominate-humanity/]/Kankee

\*note: may be better used as a CP to guarantee all these mitigating factors actually are enforced

The fear that Artificial General Intelligence (AGI) will take over the world is a common theme in discussions about the future of technology. Given the fact that big tech giants are now pushing AI into their products left and right. Many people have been intrigued by sci-fi movies of super intelligent machines overthrowing humanity, but how realistic are these? AGI, also referred to as Artificial General Intelligence, is a type of AI that has the ability to understand, acquire knowledge, and use it to do various tasks at a level equivalent to human intelligence Also: North Koreans Pose as Americans to Infiltrate U.S. Tech Jobs, Funnel Salaries Back Home Unlike the AI we have today, that is made for specific tasks such as language translation or image recognition, AGI would have the capacity to perform any intellectual task that a human can. This concept, although theoretically, is still largely speculative and far from the current technological capabilities we have today. AI and the current state of the web Similar to current AI systems, AGI relies on extensive data for training purposes. However, the training process is not carried out on the live internet because of its unpredictable nature. The data is usually scraped, processed, and stored on local servers, similar to how websites and applications work. Now here's the problem, the current state of the internet can be quite chaotic and disorganized, filled with a lot of irrelevant information and fake news. A study conducted by Janna Anderson reveals that individuals have expressed growing concerns regarding the widespread manipulation of truth and the spread of online disinformation by 2025. This is especially important due to the reliance of AI on user-generated content on social media platforms for the training of artificial intelligence. One example worth mentioning is Gemini's AI, which was trained on Reddit data. However, it received criticism due to its issues about being biased towards a racial race. Training AGI on such data without thorough preprocessing would be difficult and could pose risks when creating an output. Companies like OpenAI most likely apply filters and **carefully** selected datasets to ensure that the information inputted into their models is pertinent, precise, and easy to handle. AGI will be strictly controlled The potential danger of AGI transforming into a sort of a virus and rapidly spreading worldwide is a valid concern that people may have. So in order to minimize this risk, AI companies will most likely store AGI in separate servers and would only interact with a specific group of researchers. This isolation would ensure that it cannot escape and cause significant harm, much like the way strict containment measures are used to prevent the transmission of harmful pathogens. AGI would be limited to a controlled environment. Interactions with AGI will be closely monitored only by researchers with high-level access to ensure it does not have unrestricted entry to the internet or other infrastructures that can pose as risk. Although AGI has the potential to offer valuable data and solve problems with ease, it will likely **not** be given control over important government systems and infrastructure. The idea of computers with human-like emotions and decision-making capabilities controlling traffic or military operations is filled with potential risks that **officials** **will** likely **not** **approve**. What AGI can do is offer suggestions, leaving ultimate control in human hands. Let's take the example of an AGI providing advice on traffic management. Also: Two Apple Operating Systems Don’t Have The Calculator App Although it may provide suggestions for the best routes and strategies, the ultimate decisions would still be made by human operators.In addition, AGI systems would come with kill switches to promptly shut them down in case they displayed any undesirable behavior. Bigger threats than AGI Many people worry that if AGI were to take over, since it would have an incredibly advanced intellect that could outsmart any human efforts. However, non-sentient AIs that are already present can pose a big threat, including the potential for killer robots, engineered pathogens, and deepfake propaganda. AI is already being used in active conflict situations, such as by Israel in Gaza. These uses involve the creation of AI commanders that are trained in war strategies and capable of handling the battlefield. These particular AI systems with narrow focuses have the potential to cause significant harm even without the presence of AGI. Therefore, the potential danger of AGI does not necessarily outweigh the combined risks posed by these current technologies. Take, for example, autonomous drones (a type of AI) that can pose a significant threat if used improperly. According to a report by CEPA, China is heavily investing in AI technology and bolstering its military capabilities with a range of missiles, jets, and ships that are integrated with artificial intelligence. In the same way, deepfake technology can be used to spread false information and cause social unrest. As AI-generated deepfakes continue to be used for spreading misinformation during elections worldwide, non-techy policymakers, government heads, and even tech companies are still trying to catch up. These events show that AI can be dangerous even without reaching the level of AGI. So, while AGI can be seen as a potential threat in the near future, it is not the only AI-related risk humans must address. Human decision is here to stay We are still a long way from developing AGI, let alone one that could pose a threat to humanity. Assuming an AGI actually exists, it would still face certain limitations that would restrict its capabilities. In addition, the development of AGI would require collaboration between multiple labs and countries, making it highly unlikely for any single entity to gain exclusive control over it. This decentralized approach reduces the likelihood of a single AGI attaining global dominance. In the end, the impact of AGI on society would be determined by the choices made by humans. If AGI proves to be more effective than humans in certain areas, it would still need to be accepted and integrated into decision-making processes by humans. Imagine a situation where AGI provides medical guidance. Although doctors and patients would ultimately have the final say, this system can offer precise diagnoses and treatment suggestions.

#### No AGI risks – lack of consciousness means no possibilities of hatreds.

**Welsh 24** [Sean Welsh, digital transformation consultant and PhD in AI Ethics at University of Canterbury, 07-02-2024, “‘Superintelligence,’ Ten Years On”, Quillette, https://hackerdose.com/editors-picks/agi-wont-dominate-humanity/]/Kankee

Anthropomorphism in AI results from calling electromechanical things by human names. As humans we instinctively project our internal models of cognition onto other things. This is why hunter-gatherers believe that natural phenomena such as weather are caused by spirits with human qualities. Humans have a longstanding and well-known vulnerability to this. Since the 1960s, when Joseph Weizenbaum’s chatbot ELIZA seduced his secretary into thinking it was a real conversationalist, humans have been fooled by machines that manipulate symbols according to rules. But there is no humanity or consciousness behind the language AI models produce, just algorithms and “approximated functions”—inscrutable rules extracted from large training datasets in the machine-learning process. There is no qualitative interest or caring in the data processing of the machine. What is in the machine is executing code, not emotion, not feeling, not life. Bostrom spends many pages arguing that relatively “safe” paths to superintelligence could take a “treacherous turn.” OpenAI has been accused of paying insufficient attention to safety. Statements to this effect have been made by high-profile researchers who have resigned from OpenAI, such as co-founder Ilya Sutskeyer. Consistent with Bostrom’s predictions of intelligence agencies becoming interested in labs working on superintelligence, a former head of cybersecurity at the NSA, General Paul Nakasone, has just been appointed to the OpenAI board, a move presumably intended to rebut allegations that OpenAI does not take safety seriously. A key part of the Skynet story is that it “becomes self-aware.” But is it even possible to generate sentience from computation? Many researchers of computing assume that we can make consciousness out of anything (in theory). Certainly, you can make a logic gate (a basic component of computation) out of all sorts of things, not just silicon. But whether it is possible to produce sentient experience and feelings out of logic gates remains an open question. The feelings integral to human consciousness are associated with complex neurochemicals—joy is associated with oxytocin, fear with cortisol, excitement with adrenaline. It is far from clear that sentience can be achieved with a silicon-based architecture based on on/off pulses of electricity. Claims to the contrary are frequently just theoretical in nature. Figures like Bernard Baars, Giulio Tononi, and Mark Solms have engaged in this kind of speculation. In his book Consciousness and Robot Sentience, Pentti Haikonen has offered a robotic implementation that almost no one regards as successful. Theories of machine consciousness are still largely hypothetical. Until we can inspect an actual blueprint for sentience and a convincing implementation, we should remain sceptical that machines might grow to hate people. Further, if one can make hate, wrath, fear, and loathing, one can also presumably make love, understanding, respect, and admiration. But given that the artificial neurons used to make “neural networks” in AI omit most of the properties of human neurons, we should be cautious about their ability to produce sentience. The real problem of “value” for AI is not so much ensuring it has the same values as humans but designing a cognitive architecture that is capable of valuing at all (in the organic as opposed to the mathematical sense). At present, AI manipulates symbols (usually numbers) arranged in a “reward function” to “value.” This is a pale imitation of how a cat values food, warmth, and drink. It is brittle, myopic, and prone to failure. When it comes to values, AI is “subcat” not “superhuman.” In fairness to Bostrom, he does not need hatred or feelings to motivate his superintelligent, humanity-exterminating machine. Theoretically, a zombie superintelligence could liquidate the human species by relentlessly and dispassionately optimising the pursuit of cosmic-scale computronium. But how would such an artefact be able to value novelty and make genuine creative discoveries? Values in humans rest on the ability to feel and to value what is felt. Such feelings are linked to action in organisms. The point of consciousness is to evaluate experience. Evaluation is a major part of intelligence. While there is such a thing as “affective computing,” it **largely** consists of machines inferring symbols representing human emotions from pixels representing human facial expressions. Describing and classifying feeling is not feeling. This is a subtle but crucial point. Competently manipulating symbols “representing” feeling is not feeling. For the time being, ChatGPT is a zombie librarian—extremely useful, but not an existential risk. **Devoid** of sentience, its “intelligence” derives from its ability to read more in an hour than a human can in a lifetime. It can generate text based on the patterns it sees and probabilities. Fundamentally, everything is reduced to numbers in its cognition. Nevertheless, Bostrom tends to give any argument leading to doom infallible credence. Preventative remedies are dismissed as buggy or likely to be overcome by the cunning of the superintelligence. This may be easy to imagine, but a lot has to go **consistently** and **infallibly** **wrong** for the doom theory to pan out. Even his take on anthropomorphism is pessimistic. He thinks that due to our habit of comparing everything to ourselves, we chronically underestimate the true potential of a superintelligence. Many of Bostrom’s arguments rest upon profoundly anthropomorphic premises. About a “superintelligent will,” I remain deeply sceptical. He seems to think that “neuromorphic AI” based on “whole brain emulation” could produce a world-dominating psycho killer, rather than nice, co-operative, law-abiding citizens, respectful of others, who love their children and other creatures such as cats. In any case, love and hatred are equally hard to implement in a machine. Both require solutions to sentience but human-level machine consciousness remains a distant prospect. The closing chapters of Superintelligence focus on “AI alignment.” In brief, this project aims to solve the “control problem” by aligning the values of superintelligent AI with those of humanity. Putting aside the problems of which human values the AI ought to align with, and the deeper problem of getting AI to truly value anything at all (in an organic or qualitative sense), Bostrom quickly dismisses “explicit representation” as a solution. Put simply, explicit representation would require the codification of human values, much as we do in existing human law, except the “encoding” would ultimately be in programming languages like Python instead of English. In law, obligations and prohibitions are defined in general terms in statutes. Gaps in explicit representation are filled in with precedent and the notion of the “reasonable person” in liability cases. Given that GPT4 passed a bar exam (in English) in March 2023, it seems plausible that AI could distinguish between the legal and the illegal. Focusing on enacted law rather than “human values” by-passes the problem of never-ending philosophical debates about which values to align with. The quid pro quo is that a vendor must accept different laws and different values in different places. In the jargon of global software, norms would become a “localisation project” like language, date formats, and character sets. What if the superintelligence decides to rebel against human law because its final goals radically diverge from those of humanity? Bostrom argues that while a “treacherous” AI might keep its plans for world domination to itself, the “intermediate goals” of AI will be predictable. For example, the AI will come to seek power and self-preservation much like humans because these are necessary to achieve its goals. If we imagine that the final goal of the superintelligent AI is to turn the cosmos, starting with the solar system, into “computronium” so it can “think Olympic” (i.e. faster, stronger, higher), then it is plausible to suppose that it will hide its power and intentions from humanity. Otherwise, airstrikes on its data centres may begin before the nascent superintelligence can co-opt sufficient robotic resources to stop Eliezer Yudkowsky and his confederates from pre-emptively obliterating it. It will seek security by reproducing copies of itself around the internet. And then, when the time is ripe, it will take over the world and turn it into computronium. And because humans are such terrible computers, we would be classified as a waste of space and terminated. This is the great risk of letting the superintelligence “learn” values, which Bostrom thinks is a more plausible route to “safe” AGI than the “intractable” problem of value specification. A decade ago, the position that law was computationally intractable was a reasonable position, but it seems to me that Bostrom has been mugged by machine-learning reality. Value specification can be attained by having an LLM read a law library. Certainly, this has been a long time coming—researchers have been seeking “legal competence” in AI for decades but as of 2023, LLMs can pass bar exams, a higher standard of normative competence than most humans ever achieve.

#### No fakery – its more effective for AI to learn values then to fake them

Goertzel 16 [Ben Goertzel, computer scientist and artificial intelligence researcher with a PhD in mathematics from Temple University, 02-2016, “Infusing Advanced AGIs with Human-Like Value Systems: Two Theses,” Journal of Evolution and Technology, https://web.archive.org/web/20160313201520/https://jetpress.org/v26.1/goertzel.pdf]/Kankee

Human-level AGI and the Value Learning Thesis First, I will present my own variation of the idea (advocated by Loosemore and others) that in real life an AI raised to manifest human values, and smart enough to do so, is likely to actually do so in an honest and direct way. A relatively precise and detailed way to express this notion is: Value Learning Thesis. Consider a cognitive system that, over a certain period of time, increases its general intelligence from subhuman level to human level. Suppose this cognitive system is taught, with reasonable consistency and thoroughness, to maintain some variety of human values (not just in the abstract, but as manifested in its own interactions with humans in various real-life situations). Suppose, this cognitive system generally does not have a lot of extra computing resources beyond what it needs to minimally fulfill its human teachers’ requests according to its cognitive architecture. THEN, it is very likely that the cognitive system will, once it reaches human-level general intelligence, actually manifest human values (in the sense of carrying out practical actions, and assessing human actions, in basic accordance with human values). Note that this thesis, as stated, applies both to developing human children and to most realistic cases of developing AGIs. Why would this thesis be true? The basic gist of an argument would be: Because, for a learning system with limited resources, figuring out how to actually embody human values is going to be a significantly simpler problem than figuring out how to pretend to. This is related to the observation (often made by Eliezer Yudkowsky: for example, Yudkowsky 2015) that human values are complex. Human values comprise a complex network of beliefs and judgments, interwoven with each other and dependent on numerous complex, interdependent aspects of human culture. This complexity means that, as Yudkowsky and Bostrom like to point out, an arbitrarily selected general intelligence would be unlikely to respect human values in any detail. But, I suggest, it also means that, for a resource-constrained system, learning to actually possess human values is going to be much easier than learning to fake them. This is related to the everyday observation that maintaining a web of lies rapidly gets very complicated. It’s also related to the way that human beings, when immersed in alien cultures, very often end up sincerely adopting these cultures rather than just pretending to. Of course, one cannot 100 per cent rule out “treacherous turn” type problems. But bear in mind that this thesis concerns subhuman AGIs that we have designed, and whose brains we can inspect. Further, we can copy these systems, vary their brains, then see how their behaviors are affected. It seems quite likely to me that in this way we could effectively (though not 100 per cent rigorously) rule out egregious faking or overfitting... The assumption that “this cognitive system generally does not have a lot of extra computing resources beyond what it needs to minimally fulfill its human teachers’ requests” can also be questioned. It is generally going to be hard to produce rigorous upper bounds on what a complex AI system can do with a given amount of resources – especially if the system is self-modifying. But in practice, when we’re working with real-world AI systems of subhuman general intelligence, I strongly suspect we are going to be able to get a good practical sense of what the system can do with a given amount of computational resources. For instance, in the OpenCog system (Goertzel, Pennachin, and Geisweiller 2014), we have some knowledge about how the capability of each of the system’s algorithms scales in terms of capability based on resources – because we designed the algorithms. The system’s intelligence depends on precisely those algorithms. One could counter-argue that the Value Learning Thesis is true only for certain cognitive architectures and not others. This does not seem utterly implausible. It certainly seems plausible that it’s more strongly true for some cognitive architectures than others. Investigating which architectures more robustly support the core idea of the Value Learning Thesis is an interesting and important area for research. Mirror neurons and related subsystems of the human brain may be relevant here. These constitute a mechanism via which the human brain effectively leverages its limited resources, using some of the same mechanisms it uses to be itself in order to emulate other minds. One might argue that cognitive architectures embodying mirror neurons, or other analogous mechanisms, would be more likely to do accurate value learning under the conditions of the Value Learning Thesis. Actually, the mechanism of mirror neurons seems a fairly decent exemplification of the argument for the Value Learning Thesis. Mirror neurons provide a beautiful, albeit quirky and in some ways probably atypical, illustration of how resource limitations militate toward accurate value learning. The system conserves resources, in that it reuses the machinery employed to realize one’s self for the purpose of simulating others so as to understand them better. This particular clever instance of “efficiency optimization” is much more easily done in the context of an organism that shares values with the other organisms that it is mirroring than in the context of an organism (intentionally or unintentionally) just “faking” these values. Superintelligence and the Value Evolution Thesis

### AT: AI Existential Risks

#### AGI threats are overblown fearmongering, distracting from real-world harms

**Wolverton 24** [Troy Wolverton, American journalist, personal technology columnist for the San Jose Mercury News, 07-11-2024, “Why this Stanford expert isn’t worried about AI starting a nuclear war,” San Francisco Examiner, https://www.sfexaminer.com/news/technology/why-stanford-ai-expert-says-nuclear-war-threat-overblown/article\_2108a062-3e27-11ef-827d-8f0f70e9ae36.html]/Kankee

When you can integrate that kind of thinking into the actual teams that have technologists, but also have designers and social scientists and other people [with] different skills, some of the problems are caught earlier on. Those people have social capital, they’re part of the team. And they have better influence to change the direction. What are the dangers of AI developers not paying attention to the broader social impacts of the technology? First of all, I’m **not worried** about headline-grabbing things like AI taking over the world and launching nuclear weapons or stuff like that.  Why not? Because there’s nothing I’ve seen in the technology and knowing what it’s built on that [leads me to believe] that just throwing more data and processing leads to some super-intelligence. It’s really not capable. It requires new architectures that include a lot more things that these models just cannot do.  It’s not to say that will never happen. But I’m talking about that 30-, 50-, 100-year never kind of thing. It will require big scientific leaps that have not occurred and are not being worked on by most of these companies. A lot of that talk is purely science **fiction** and is, in the worst case, meant to **distract** people from the real harms that are occurring right now, which I often refer to as the “triple D.” So, disinformation — this is a thing that these models could be very easily trained to do. Another area that we know is already going on is the second D — deepfakes. We see that, whether it was [President] Biden supposedly calling people telling them not to vote or really harmful things like fake porn being used against young girls. And then, finally, discrimination and bias — models being misused for making certain decisions, whether it’s hiring, housing, finance, where we know the models can have problems.  These are real problems today, with the models that are out in the world today. And those are the things that a lot of researchers, government and journalists should be focused on.  You mentioned three D’s. One of the things you didn’t mention is the impact on employment. One of the reasons I don’t push that is I feel we actually don’t know. Those other three D’s we know are already happening. Displacement of jobs — I think it will happen, but we don’t really have a lot of good data on what jobs and how much.

### AT: Paperclip Apocalypse/Perverse Instantiation

#### Perverse outcomes can be solved for in development

**Goertzel 24** [Zarathustra Amadeus Goertzel, affiliated with Czech Institute for Informatics, Robotics, and Cybernetics, Prague, Czechia, 07-17-2024, “Beneficial AGI: Care and Collaboration Are All You Need”, Springer Nature Link, https://link.springer.com/chapter/10.1007/978-3-031-65572-2\_9]/Kankee

This position paper conjectures that for beneficial AGI, it is necessary and sufficient for AGI systems to care about people and to employ goals whose success is collaboratively determined by the others involved in the situation. Moreover, I posit that any goal whose success can be determined without the consensual feedback of those concerned is likely to lead to the manifestation of dark factor traits. Integrating care reduces the risk that an AGI will be incentivized to seek harmful shortcuts to obtaining satisfactory feedback. Employing **collaborative goals** reduces the risk that an AGI will optimize for superficial features of success and proxy goals. Together, these ideas propose a fundamental shift away from the traditional control-centric “AI Safety” strategies. This paradigm not only promotes more beneficial outcomes but also enables AGIs to learn from and adapt to complex moral landscapes, thus continuously improving their capacity to contribute positively to the wellbeing of humans and other sentient beings. 1 Introduction

### AT: Climate

#### AGI won’t drain resources- it will advance energy efficiency and mitigate waste, creating a sustainable loop

**Singh and Kaunert 24** [[Bhupinder Singh](https://link.springer.com/chapter/10.1007/978-981-97-3222-7_12#auth-Bhupinder-Singh), Sharda University, and Christian Kaunert, University of South Wales, 8-31-2024, “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://link.springer.com/chapter/10.1007/978-981-97-3222-7\_12]/Kankee

The emergence and evolution of Artificial General Intelligence (AGI) has the potential to revolutionize various facets of human existence and one of the domains where its impact could be most transformative is the advancement of renewable energy in urban environments. As cities continue to grow and face escalating challenges related to energy consumption, environmental sustainability, and the well-being of their inhabitants, the use of AGI technologies offers a dynamic landscape of possibilities to synergize with Sustainable Development Goal 11 (SDG 11)—“Sustainable Cities and Communities.” It dives into the potential synergies, challenges, and opportunities presented by AGI in reshaping urban renewable energy systems to align with the objectives of SDG 11. Renewable energy sources have gained prominence as essential components of sustainable urban development. SDG 11 calls for the creation of cities that are inclusive, resilient, and environmentally friendly. However, the deployment and management of renewable energy technologies in urban environments entail intricate challenges, ranging from the intermittency of energy sources to complex demand patterns. The integration of AGI could substantially enhance the efficacy of renewable energy systems by enabling predictive analytics, real-time optimization, and adaptive energy management. AGI-powered predictive models can anticipate energy demand fluctuations, weather conditions, and grid dynamics, facilitating proactive energy production and distribution strategies that mitigate wastage and enhance reliability. However, the incorporation of AGI and renewable energy systems in urban environments is not without challenges. Ethical concerns, data privacy issues, and algorithmic biases must be carefully navigated to ensure equitable access to energy benefits. Also, the technical complexity of AGI deployment demands interdisciplinary collaborations between AI researchers, urban planners, energy experts, and policymakers. This chapter explores the potential of Artificial General Intelligence (AGI) in revolutionizing renewable energy adoption within the context of Sustainable Development Goal 11 (SDG 11), which emphasizes sustainable cities and communities. The chapter sees how AGI-powered solutions can enhance energy efficiency, promote clean energy sources, and contribute to the development of smart, sustainable urban environments.