Internet Governance Project

IGP White Paper #11 [May 2024]

# The Myth of AGI

## How the illusion of Artificial General Intelligence distorts and distracts digital governance

**By Milton Mueller**

Executive Summary

The claim that Artificial General Intelligence (AGI) poses a risk of human extinction is largely responsible for the urgency surrounding AI governance. This paper reviews and critically evaluates the AGI-related literature in computer science, economics and philosophy to understand the assumptions and logic underlying claims that AI can threaten human survival. The review identifies three inter-related fallacies underlying AGI doomer scenarios**:** a) the idea that a machine can have a "general intelligence;" b) anthropomorphism, or the attribution of autonomous goals, desires and self-preservation motives to human-built machines; and c) the assumption that the superior calculating intelligence of an AGI will give it unlimited power over physical resources and social institutions. The paper characterizes these assumptions as unrealistic and exposes the lack of logic and empirical evidence in the doomer scenarios. Evaluating the AGI construct is important from a public policy perspective because of the myth's enormous influence on the way governments, industry and the public approach digital governance. The idea of an all-powerful autonomous AGI misdirects policy interventions toward precautionary regulation of the design or production of all AI applications, while diverting our attention from more mundane, yet realistic risks posed by specific AI uses and users. The myth of existential risk also encourages governments to attempt to assert control over the entire digital ecosystem in ways that stifle competition and innovation and centralize power.

## Introduction

In March 2023, more than 1,000 technology business leaders, researchers and intellectuals signed an open letter urging a moratorium on the development of artificial intelligence, claiming that it posed "profound risks to society and humanity." Two months later, an open letter signed by more than 350 executives, researchers and engineers claimed that artificial intelligence posed a "risk of human extinction" and urged us to make "mitigating that risk a global priority." (Center for AI Safety, 2023).

In response, the world's governments and various global governance institutions leapt into action, or at least provided the appearance of such, to meet the alleged threats. The G7 started an AI Hiroshima process, the very name invoking the nuclear destruction that ended World War 2. The US Congress held hearings in which industry incumbents reinforced the warnings.[1] The U.S. president published an Executive Order (EO) claiming to take "the most sweeping actions ever taken to protect Americans from the potential risks of AI systems." The EU passed what it claimed was "the world's first comprehensive AI law" late in the year. Many of their regulations, however, were anticipated by more focused actions taken by the Peoples Republic of China, (Sheehan, 2023) and China was not having a conversation about the risk of human extinction. Its government was focused laser-like on taming and controlling AI's ability to strengthen or weaken governmental control of public expression. (Zhang, 2024).

> "…probably Man will construct the deus ex machina in his own image." I.J. Good. Speculations concerning the first ultra-intelligent machine, *Advances in Computers*, 6, 1965.

What accounts for the sudden, apocalyptic turn in the discourse about AI? Critiques of possible harms from AI applications, ranging from general concerns about bias (Friedman & Nissenbaum, 1996) errors in facial recognition (Leslie, 2020), and autonomous vehicle safety (Koopman & Wagner, 2017), have been raised for many years. The claim that AGI technology threatens all of society with extinction, on the other hand, is a claim that, while not new, was not part of the mainstream until recently. How did we get to a place where the progress of computing technology is routinely paired with a risk of societal extermination?

The answer is that almost from the moment computer technology was invented, philosophers, science fiction writers and some developers invented a vision of Artificial

[1] An oversight hearing to examine artificial intelligence, focusing on principles for regulation. Senate Judiciary Subcommittee on Privacy, Technology, and the Law. July 25, 2023. "Oversight of A.I.: Legislating on Artificial Intelligence," Senate Judiciary Subcommittee on Privacy, Technology, and the Law, September 12, 2023.

General Intelligence (AGI). AGI is defined as a superhuman intelligence that, somehow, gains the ability to act independently of human instructions. The label AGI is relatively new; earlier literature used different labels for the same idea: superintelligence, ultra-intelligent machines, etc. As far back as 1965, an Oxford computer scientist defined an ultra-intelligent machine as “a machine that can far surpass all the intellectual activities of any man however clever” (Good, 1965). Later ruminations speculated that AGI would be the product of a “singularity” in which machines gained self-consciousness and autonomy, posing the risk of a "runaway reaction" of self-improvement cycles that would make it so powerful so rapidly that humans would not be able to catch their breath before their very existence was threatened. (Kurzweil, 1999; Bostrom, 2003) I.J. Good (1965, p. 33) gave this vision a slightly more optimistic spin: “The first ultra-intelligent machine,” he wrote, “is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”

Almost from the beginning of machine computing, then, the promoters of this messianic vision have assumed that advanced machine intelligence would also be autonomous and have unconstrained powers over the physical world, including the power to eliminate or destroy human society.

The early dialogue about an AGI was a relatively playful one, sitting somewhere on the border of philosophy and science fiction. Discussion in the 1990s was confined to obscure chat groups and email lists populated by AI developers with grand, sometimes creepy, philosophical notions. (Murgia and Thornhill, 2023) The recent, impressive success of large language models in generating natural language responses to human queries, however, convinced many developers that we are on the verge of creating an AGI – or may have already done so. With the vision of AGI already defined by leading thinkers as an autonomous, all-powerful and possibly malevolent being, the intellectual groundwork was laid for today’s panic.

This paper addresses these fears head on. It carefully examines the scholarly literature on AGI. Its review of this literature identifies and critically examines the assumptions underlying autonomous AGI scenarios. It is shown that three inter-related assumptions underlie the "existential risk to humans" scenario: 1) the idea that a human-made machine can have a “general intelligence;” 2) an assumption that machines will acquire their own desires and utility functions; and 3) the assumption that superior machine intelligence will give AGIs unconstrained control of social institutions and material resources. All three of these assumptions are exposed as unsupported.

Section 1 of this paper attacks the validity of the AGI construct itself. It demonstrates the inability of scientists and philosophers to provide a meaningful definition of what “general” intelligence is. Section 2 takes up the question of autonomy – whether an AGI can set goals for itself independently of humans and break free of human control. I review recent computer science literature purporting to provide a scientific basis for this possibility and show that the assumptions underpinning this work have no empirical grounding and are based on circular reasoning. Section 3 takes up the problem of power and physicality. It shows that the "catastrophic risk" case rests on yet another false assumption, namely that superior machine intelligence automatically yields unlimited physical power. The last section discusses the governance implications of basing public policy on the specter of an AGI. It shows how the AGI myth encourages policy responses that are both impractical and, should they be attempted, authoritarian in their effects on human use of ICT. It ends with a call for humans to take responsibility for what they do with technology, rather than offloading it to allegedly autonomous machines.

## 'General' Intelligence

What is an AGI? Most definitions start by differentiating artificial general intelligence (AGI) from artificial narrow intelligence (ANI). This conceptual move is necessary because computerized intelligence already vastly exceeds the capacity of humans wherever it has been successfully applied. AI applications can beat humans at board games. Computers have been calculating solutions to mathematical or scientific problems faster and more accurately than humans for decades. Computers can perform data analytics better than humans. No one has asserted that these applications pose a risk of human extinction, however.

The societal risk of an AGI, therefore, cannot be attributed to the mere fact that intelligent machines outperform humans in any given task. Instead, the AGI literature labels all existing AI applications as "narrow." Chatbots, language translation, chess playing, spam filtering, social media recommendation algorithms, medical diagnosis and facial recognition, to name a few, are all considered ANI. In the words of McLean et al (2021), "…an ANI's intelligence is task specific (or narrow) and cannot transfer to other domains with unknown and uncertain environments in which they have not been trained." They go on to say,

> "…an AGI would possess a different level of intelligence, which has previously been defined as an agent's ability to achieve goals in a wide range of environments, and the ability to achieve complex goals in complex environments." (McLean et al, 2021)

The godfather of the AGI construct, philosopher Nick Bostrom (2003; 2014) defines AGI as a "superintelligence" and (echoing Good 40 years earlier), claims that it could "vastly outperform the best human brains in practically every field, including scientific creativity, general wisdom, and social skills."

Two key questions are always overlooked in these discussions: what goals would this "general" intelligence have, and where would they come from? Progress in real-world machine learning has always come from humans programming and training AI applications to execute specific tasks. The ability of Large Language Models (LLMs) to generate sentences and translate languages, for example, came after years of conceptual modeling of grammar and the input of vast stores of digitized texts. Facial and image recognition applications have been trained on millions of digitized images and programmed to match individual faces to identities. Other AI applications are based on complex algorithms that are constantly updated and trained to do what humans want them to do. Generally, all these applications perform better the more well-defined their goals are. ChatGPT and other applications with impressive text-generation capabilities, for example, are notoriously bad at simple arithmetic. The powerful neural networks of these LLMs cannot do math as well as a 30-year-old pocket calculator.

The notion of an AGI, therefore, is a qualitative departure from everything we currently know about machine intelligence. Instead of learning to do something better than humans, an AGI is supposed to be a single application that can learn to do anything and everything better than humans. This, however, is a meaningless concept, an emperor with no clothes. In what may be the only definitional discussion of AGI that faces this deep conceptual problem, Phillips (2017) wrote:

> "The purview of Artificial General Intelligence (AGI) is the development of theories, models and techniques for the endowment of machines with intellectual capabilities that generalize to a variety of novel situations. This characterization, however, [begs] important questions about what we mean by *intelligence* and *generalize*. In the absence of precise criteria, researchers look to the archetype of general intelligence, human cognition."

Here we encounter the unseen problem at the very center of the AGI construct. There is no scientific definition of what "general intelligence" is. Our only model for general-purpose intelligence is the human being, and an AGI's intelligence is said to be both general-purpose like human cognition, yet vastly superior to humans, at the same time. To invoke Laurie Anderson's satiric description of paradise, an AGI's intelligence is exactly like a human's - only much, much better.[2]

There is no scientific definition of "general intelligence." Our only model for general-purpose intelligence is the human being, and an AGI's intelligence is said to be both general-purpose like human's, yet vastly superior to humans at the same time.

Aside from being a highly stretched analogy, the use of human intelligence as the "archetype of general intelligence," contains a massive oversight. The "generality" of human cognition is rooted in our status as a living organism whose survival strategies involve tool use, language and social cooperation. In other words, it is life – the human being's imperative for survival, sustenance, shelter and reproduction - that provides human intelligence with its objectives and its evolutionary trajectory. Humans have learned to adapt to novel situations

[2] "Paradise is exactly like where you are right now... only much, much better." Anderson, Laurie, "Language is a virus (from Outer Space)" https://mojim.com/usy102988x9x9.htm

because it enhances their survival as a species. Intelligence serves life, not the other way around.

Yet computers are not alive. Giving them more computing power, building more complex neural networks, and feeding them more data will not by itself make them living beings. Life requires the ability to act autonomously to find the resources needed to sustain operations, and the ability to self-replicate. Computers cannot do that yet.[3] No existing computing machines can solve problems they have not been given. They must be told what objectives to pursue and must be trained to pursue them. (I will address the question of whether AGI will develop its own goals later.) In other words, humans make computing machines "more intelligent" by improving their ability to pursue specific objectives defined by their human creators. Ergo, a "general intelligence" attributed to a man-made computing machine is really an oxymoron.

Advocates of AGI may try to fall back on the concept of "meta-learning" to salvage their pursuit of the AGI vision. Meta-learning refers to AI applications which program machines to learn to learn. Do self-improving AI systems bring us closer to the promise of AGI? Some computer scientists suggest that it can: "Meta-learning provides a promising paradigm that allows AI systems to learn from prior experiences and generalize that knowledge to new and unseen tasks." (Orike & Ene, 2023) Yet, when one looks at the actual research on self-improving AI one finds that its generalizations are always in specific domains.[4] Meta-learning simply reduces reliance on training data by substituting self-programmed recognition of patterns for massive volumes of data. It does not eliminate the need for training, however, and its “generality” is restricted to the specific knowledge or action domain for which the machine was constructed. The machine's objective is still a “narrow” one given to it by humans.[5] Progress in computing may involve making AI decisions more adaptable and their analytical capabilities more generally applicable to its defined purposes, but it does not give machines life, nor does it give them autonomy.[6]

[3] In his 1945 theory of automata (von Neumann, 1945), von Nuemann tried to model a machine that could self-replicate, but as of yet that theory has not been put into practice.

[4] "In computer vision, meta-learning can enhance the ability of AI systems to recognize objects, segment images, or detect anomalies with minimal training examples. In natural language processing, meta-learning can facilitate language understanding, machine translation, and question-answering tasks with limited labelled data. In robotics, meta-learning enables robots to adapt quickly to new environments, tasks, and interactions by leveraging prior experiences." In other words, the self-improving AI applications are making the machines more efficient at specific tasks, not a "general" AI modeled after human life. (Orike & Ene, 2023, p 13)

[5] For that matter, why would anyone build a machine that performed tasks that were not defined or controlled by its maker?

[6] Of course, humans *can* create self-sustaining forms of life, but that does not require making them more intelligent, much less super-intelligent. Many highly successful forms of life – rats, cockroaches, plankton – are less intelligent than humans. Humans do need to worry about the risks of creating alternate life forms that are infectious or dangerous. Yet a robot capable of replicating and surviving would have to be built and programmed to do so by humans, and would not have to be “superintelligent” or anything close to it. It would just need to know how to replicate and survive. True, if a machine turned into a life form it would be more likely to survive if it was

As an example, consider the case of an Autonomous Vehicle (AV) designed to navigate vehicle and pedestrian traffic in complex urban environments. The AV may use meta-learning to find new and more effective ways to recognize and avoid collisions, or more efficient ways to get to its destination. But we should not expect it to come up with new ways of recognizing the causes of cancer or to engage in remote eye surgery. It would not be physically or computationally equipped to do those things. And why would its builders, investors and owners want it to do these "new and unseen" tasks?

The fallacy underlying the AGI construct can be clarified by asking, what is a "general-purpose" machine? Computers are, after all, computing machines. Any workable machine that one can think of has a delimited purpose or set of purposes given to it by its designer/builder, and the more well-defined and specific its purpose(s), the more efficient it is likely to be. Machines can combine multiple functions, but the concept of a "general-purpose machine" is literally meaningless, both semantically and operationally.[7] The claim that we can build a machine with generalized intelligence is logically equivalent to a claim that we can build a single machine that does everything. It makes no sense.

## 2. Autonomy and Machine Evolution

It should be clear from the last section that when computer scientists talk about creating an AGI, they are really talking about creating life.

The inability of philosophers and computer scientists to provide a scientifically meaningful definition of a man-made "general intelligence" creates the space into which religious visions of computers coming alive can be projected. As we trace its next steps, we will see the claim that an AGI is an existential threat become increasingly anthropomorphic, attributing life, motives and supreme power to computing machines.

A key step in the existential threat argument is that an AGI is assumed to have values of its own that are not given to it by humans. In the words of Bostrom (2003), "general superintelligence would be capable of independent initiative and of making its own plans and may therefore be more appropriately thought of as an autonomous agent." How would a computing machine achieve this autonomy? Bostrom and the earlier literature merely fantasize about it. In Bostrom's own words, his description of superintelligence "leaves open how the superintelligence is implemented…" It is purely a thought experiment.

Some of the more recent literature in computer science, however, attempts to show how computing machinery might become autonomous. Human intelligence evolved from nature;

able to solve a broader range of problems thrown at it by its environment, but there is no reason to assume it would need "superintelligence" to do that. The "generality" and scope of its intelligence would be bounded by the survival and reproduction requirements dictated by its environment.

[7] Automobiles are machines for transportation, but while technical innovations may allow us to use the same transport machine to drive on roads, fly, sail over oceans, and perhaps even embark on space travel, we cannot make it also a rug cleaner, a manufacturer of homes, a stock trading platform and a drug dispenser.

mightn't AGI evolve from computers? These researchers attempt to show how an AI application might acquire autonomy and a will to survive through known features of deep learning, reward structures, and the application of game-theoretic models. The progression is based on three arguments: "the alignment problem;" the notion of "AI drives" (Omohundro, 2007; Shulman, 2010), and a belief that these drives can prevent humans from disabling or controlling the machine – the so-called "off-switch problem." (Hadfield-Menell et al, 2008; Sotala & Yampolskiy, 2015).

The alignment problem is defined as "the challenge of ensuring that AI systems pursue goals that match human values or interests." (Ngo et al., 2023; Russell, 2019; Gabriel, 2020) The "AI drive" work asserts that "goal-seeking systems will necessarily begin to model their own operation and to improve themselves" in ways that give them their own motives. (Omohundro, 2018) The off-switch problem is a simple game theory model of why a machine might resist being turned off. Taken together, these arguments try to mount the case that machines equipped with advanced artificial intelligence could evolve into a life-form – and a potentially dangerous one at that.

Insofar as there is a coherent argument here, it indicates two possible paths toward an AGI that is a threat to human control. The first is that the machines will evolve into an autonomous AGI (Bullock, Mckernon, & Dicarlo, 2023); the other is that AI training and development would result in a deviation-amplifying feedback process (Ngo, et al, 2023) through which the machines acquire their own purposes and autonomy from human purposes.

The machine evolution argument can be readily dismissed. Machines do not evolve. Of course, technology does change over time in ways that superficially resemble "evolution," but a proper application of evolutionary concepts corrects that perception. The machine itself isn't evolving – the changes in technology are produced by humans responding to markets and other human social systems. Evolution in the Darwinian sense requires self-replication, mutation and natural selection over many generational cycles. The replication of machines comes from human manufacturing processes, not from self-replication. Nor does mutation – changes in the design of machines – come from the machines. They come from human decisions in response to societal pressures for efficiency, innovation and safety. The selection process – that is, which machines continue to be produced and which become obsolete – is also governed by human decisions. The choices are made by people in competitive markets and/or political institutions. Hence, the machines themselves do not evolve, industries and socio-technical systems do. Human decisions, by individuals or in aggregate, control each step of the process. If this is true, we cannot speak of "machine evolution." For machines to evolve, they would first have to become alive, which means they would gain the ability to self-support and self-replicate, which they currently cannot do.

So, the autonomous AGI scenario must be based, at least initially, on a cybernetic process, in which human efforts to produce AI generate a deviation-amplifying feedback loop that not only makes machines more intelligent but also gives them their own goals and the ability replicate themselves or "survive" without human consent.

The alignment problem literature starts by making a plausible case that there can be gaps between the objectives that humans hope to reinforce with their AI models and training, and the actual behavior the machine learns. This moderately interesting fact about AI models is then twisted into two unwarranted conclusions: 1) that these gaps will progressively enlarge until the machine develops internally defined objectives that are unrelated to the objectives programmed by their trainers, and 2) that humans will not notice these gaps and/or will be unable to correct them.

Because of their unfamiliarity with social studies, AI doomers in computer science fail to realize that an alignment problem is not unique to AI training. Similar uncertainties and misspecifications characterize all education, legislation, and contract negotiations. We may think we are training children to behave in a certain way in schools and families, but they may draw different conclusions and behave very differently.[8] We may think that a law structures human behavior in ways that the government wants, but there are often unintended or even perverse effects as humans, who really are intelligent and autonomous agents, find ways to exploit the new rules. We may think that a contract sets out an agreement that satisfies both parties, but contingencies and problems may arise that are not clearly covered by the contract.[9] An alignment problem exists in all forms of human-human and machine-human interactions, because humans cannot always specify with perfect clarity the objectives they want an external party, whether human or machine, to pursue. Generally, we overcome this problem through ongoing adjustments based on trial and error, theory and learning; i.e., through deviation-reducing feedback. We also address it through institutionalization – by adopting rules and conventions designed to narrow the gap between expectations and behavior.

To move from a garden variety alignment gap to an existential threat to humanity, the doomer Computer Scientists must argue that minor deviations between the intentions of human designers and the reward and training structures of the machines will be progressively amplified until the machine(s) pursue objectives that humans have not given them and do not want. They must also assume that humans cannot intervene in this process to correct those deviations - the amplification must continue uninterrupted until it takes a dangerous turn.

[8] By the AI safety logic, an alignment problem in education might create a risk of human extinction by producing a Stalin or Mao or Hitler, so this should lead us to "pause" all educational activities.

[9] This is called the problem of incomplete contracts in institutional theory. Indeed, institutional theory often uses the same word, *alignment,* to describe how social structures harness or channel the incentives of actors with divergent interests into cooperative, socially beneficial goals. (Spiller, 2009; Foss 1996)

It is possible for alignment gaps to develop, persist and perhaps even widen slightly over time. To conclude that advanced AI applications might develop their own goals, however, doomers must also assume that humans will not be able to see the gaps happening and make *any* corrections at *any* time.

Ngo et al (2023), for example, argue that AGI's trained through reinforcement learning from human feedback (RLHF) "will likely learn to plan towards misaligned internally represented goals that generalize beyond the RLHF fine-tuning distribution." Somehow, the original objective given by human developers and trainers is subordinated to endogenously developed ones of its own. The doomers even posit that the machine will lie to humans to cover up its deviance and manipulate its reward system to gain more power to pursue its own ends. In other words, the machine is alive, wants to be free of human control, and will acquire the tools and resources to become so. The threat of human extinction from AGI simply fails to materialize unless they take the anthropomorphic leap.

Yet, computer science literature never provides a model of a cybernetic process that would produce these results. Indeed, none of the papers making assertions like this demonstrate formally that this could happen, nor empirically that anything close to it has happened. Nor is there any statistical or mathematical analysis to estimate how "likely" this would be. In fact, it is presented as another imagined possibility, based on anecdotes about minor alignment gaps found in AI labs.[10] They fall far short of showing any inherent progression toward “misaligned, internally represented goals.”

It is possible for alignment gaps to develop, persist and perhaps even widen to some degree over time. It depends on how the system is governed. To conclude that advanced AI applications might at some point threaten human life, however, the AI doomers must also assume that humans will not be able to see the gaps happening and make any corrections at any time. In other words, humans are assumed to be unable to intervene in the cybernetic process (or the process of machine evolution).

If one tries to concretize the out-of-control machine evolution or deviation-amplifying cybernetic process arguments, their absurdity becomes apparent. Imagine an autonomous vehicle service – an Uber or Lyft without drivers – operating in the complex urban environment of Amsterdam, The Netherlands. Its objective is to charge profitable fees for a transport service that accepts pickup points from customers and delivers them to a destination they specify. And before it can go into operation it must satisfy basic vehicle safety standards

[10] The fact that these alignment gaps were noticed and discussed in the literature is a self-refutation of their argument that these gaps would not be noticed and would get worse and become dangerous.

from the local government. The system requires massive capital investment in vehicles, the development of powerful AI models and extensive training, so its human owners are not going to be casual about supervising its operation. Let's say the system eventually is implemented in production and succeeds in gaining the trust and business of millions of customers by performing its function well (because if it did not function well enough initially, it would never be operated long enough to create a widening alignment gap).

But then, according to the AGI doomers, small alignment problems and reward hacking tactics "would likely" build up. Eventually, a customer who wanted to go from the Amsterdam train station to the Rijksmuseum is instead delivered to Tilburg, or maybe to Brussels or (why not?) Northern Africa. Or it might pursue an even wilder deviation and use all its vehicles to initiate an electric vehicle race through the streets of Amsterdam. The doomers must argue that such a massive deviation would arise instantly and without any warning, that no one would notice the smaller misalignments that led to this big problem, and that once it happened, there would be no human ability to withdraw the vehicles from service, shut down the computers, modify the algorithms, sue the owners, etc.

### 3. The Off-Switch Problem

It gets sillier. The misalignment between human intent, human control, and the actions and objectives of the machines, they claim, will somehow create a powerful urge for self-preservation in the machine. The justification for this leap of logic is a game theoretic model known as the "off switch problem" (Hadfield-Menell, D., et al., 2008; Sotala & Yampolskiy, 2015). This game is based on the idea that machines that have been programmed to pursue a certain objective will realize that they cannot pursue that objective if they are turned off, or "dead." As Sotala & Yampolskiy put it,

> "...many formulations of rational agents create strong incentives for self-preservation ... a rational agent will maximize expected utility and cannot achieve whatever objective it has been given if it is dead."

To refer to our autonomous vehicle example, an AI application programmed to move customers in Amsterdam from point A to point B will realize it cannot do that if it is turned off. So it will act to stop anyone from disabling it. Note the key, unstated assumption: the machine has the *physical power* to prevent itself from being turned off by any external agent. Note the obvious oversight: the doomers recognize the existence of a utility function that guides the machine's action but overlook the fact that this utility function has been given to the machine by (some) humans and serves their ends. Further, the AI doomers are serving up two contradictory stories: they say the AGI is so devoted to its AV transportation function that it will fight human intervention to turn it off so it can keep doing it, while at the same time saying that the AI governing the system will feel no inhibitions about radically departing from that function to pursue ends of its own.

There are many appeals to utility functions in this literature, giving readers the impression that the writers are well-versed in economic theory. They are not. Utility functions are subjective to living individuals. You must be alive to have preferences. Machines get their preferences from humans. And if they come from humans, a badly specified utility function that leads to bad behavior can be replaced after humans notice that it is producing bad results.

Furthermore, utility functions measure benefits at the margin. One thing is traded off for another until their relative proportions reach an optimizing equilibrium. More and more of the same thing is never optimal. Yet the doomer scenarios ignore this. To draw on a famous example, if an AGI's objective function is to produce paper clips, the doomers claim that its devotion to this objective would be so strong that it just might turn the entire world into paper clips. That argument loses sight of the fact that the marginal benefit of producing another paper clip gradually declines as supply increases. Likewise, the marginal cost of supplying the inputs needed to make paper clips would increase rapidly as paper clip production crowded out other possible uses. As it attempted to consume more and more of the world's resources, the paper clip AGI would see the price of the raw materials and energy inputs rise to the point where it no longer made sense to produce another paper clip (much less attempt to turn humans – who are not the most efficient source of raw material – into paper clips). Indeed, if its utility function gave it an instinct for self-preservation, as the AGI doomers believe, then surely it would realize that continued consumption of resources would threaten its own functioning (it would not be able to produce more paper clips if it turned itself into one, would it?).

The AI doomers often appeal to economic theory, but their understanding of the objective function seems to miss the most basic insights of economics. Even if a machine's utility function was so crude as to not incorporate the basic constraints of marginal utility, where would the AGI get the money to keep buying all those inputs as the prices rose?

Once again, the doomers have to make a fantastic leap of reasoning. They say the AGI would be able to overpower all other agents to pursue its objective. The machine would be able to steal or appropriate whatever it needs – because its utility function told it to. Appealing to economic theory, Sotala and Yampolskiy, 2015 say:

> "AGI systems which follow rational economic theory will then exhibit tendencies toward behaviors such as self-replicating, breaking into other machines and acquiring resources without regard for anyone else□s safety."

Not only has the AGI become a criminal homo economicus with its own subjective preferences, but its pursuit of an objective with no limit (contradicting the marginalism of economic theory) will magically overpower any internal or external constraints human place on it. The machine is assumed to have perfect knowledge of the way changes in its programming or physical composition will affect its activities in the future (another assumption totally at odds with economic theory). And along with its "drives" comes unlimited power – the power to rebuild itself, "hire outside agencies," cheat, steal and destroy. (Omohundro, 2007, 2008)

If this is a real possibility, why haven't chess-playing AIs evolved the ability to cheat? Why don't they make their Queen, which was taken by the opponent two moves earlier, suddenly reappear on the board? It is, after all, strongly motivated to win the game, and another queen would certainly help. The fact that the rules of chess are programmed constraints doesn't matter in AGI doom scenarios. In the words of Omohundro (2018), "it just alters the landscape within which the system makes its choices. It doesn't change the fact that there are changes which would improve its future ability to meet its goals." It might even figure out dangerous ways to circumvent the rules; e.g. killing off anyone who objected to its cheating moves, or mentally disabling its best human (or AI?) opponents before they even sat down at the chess board. In other words, in doomer scenarios the goal always overrides the rules, the programming, the physical constraints.

The discussion here makes it clear that the anthropomorphic leap, which is questionable enough, does not by itself create a catastrophic risk. That requires yet another irrational leap, the omnipotence leap. The omnipotence leap says that after setting its own goals and establishing itself as an autonomous, self-replicating life form, the AGI also has unlimited power.  The omnipotence leap is discussed next.

## 4. Physicality

Computer scientists do not think in terms of social institutions and material resources. They tend to think of the off switch as a binary logic gate. It is either off or on, and to toggle between the two you only need to send a signal. Survival in the real world, however, is not just a matter of manipulating symbols. It involves physicality.

If an AGI is going to win a fight with humans over the off switch, it must have effectors. An effector in cybernetics is a mechanism capable of acting on the physical world in response to instructions from a controller. An AGI capable of threatening humans with extinction must be capable of much more than calculation, information processing and messaging. It must be a cyber-physical system (CPS) with physical appendages or weapons, and sufficient energy resources to operate them. To pose a credible threat to all of humanity, in fact, an autonomous AGI cannot just be a single, isolated CPS like a Reaper drone. Its physical effector systems would require massive scale, huge force multipliers, almost unlimited physical inputs, and control of the many social systems (money, power sources, weapons, communications) required to deliver the required inputs and impose its will.

The existential threat imaginary assumes that superior calculating capability gives an AGI omniscient physical power unchecked by any social and physical constraints.

Here we find the deepest flaw in the AGI autonomy argument. Even supposing that one or more machines managed to generate internally developed goals not controlled by humans, the existential threat imaginary must go on to assume that the AGI is ultra-powerful and unchecked by any social and physical constraints. Super intelligence somehow allows a digital information system to consume unlimited quantities of electrical power and other scarce resources to muscle out all its competitors. Its effectors are assumed to have the power to prevent humans – whether individuals or organized armies – from disconnecting its power source, turning it off, improving or correcting its algorithms or utility functions, or destroying it. It can steal or pay for all the inputs it needs.

No one ever explains how this could happen. They just assume that computing intelligence alone enables the machine to overcome all external physical constraints on whatever behaviors it wishes to enact. The simple off-switch game model starts by assuming that the machine has the power to prevent itself from being turned off (Wängberg, Böörs, et al, 2017), and then models whether its utility function would make it decide to allow that or not.

We see this fallacy most clearly in Bostrom's description of superintelligence. He attributes practically divine powers to it, saying, "It is hard to think of any problem that a superintelligence could not either solve or at least help us solve. Disease, poverty, environmental destruction, unnecessary suffering of all kinds: these are things that a superintelligence equipped with advanced nanotechnology would be capable of eliminating." Note the idealist assumption that solving these problems requires only "intelligence;" it would not require the construction of physical equipment, enormous amounts of labor and energy, or large re-allocations of capital away from other purposes. A disembodied Mind could do it all. Look more closely, however, and we see that in an aside Bostrom slips in a reference to the AGI's physicality – it is "equipped with advanced nanotechnology." The idealist assumption is abandoned. Once again, the AGI acolytes assume away the most significant problem – how does it do that? - by saying that an AGI's superpowers come from effectors deploying advanced technologies that do not exist yet but will magically appear once it arrives.

Recall Bostrom's previously mentioned climbdown from his notion of superintelligence: the concept "leaves open how the superintelligence is implemented…" And therein lies the rub. The risk of machine intelligence getting out of control and threatening human society cannot be assessed without specifying just that: how it is implemented, materially. Who built it? Who funds it? How are its components geographically distributed? What are its power sources? How is it physically connected to society's communications, power and transport infrastructures? How many competing or countervailing agents exist, whether human or machine? All AGI doom scenarios ignore these questions.[11]

Now imagine humans struggling with an autonomous AGI that is damaging their world or seizing control of it. They are engaged in active combat with it. The AI doomers are positing

[11] Pinker, (2019), makes a similar point.

that humans will always lose. In short, they say an AGI will become an alternative life form that pursues its own objectives, avoids its own death, replicates itself, and can successfully resist all efforts by the planet's dominant life form - humans - to correct it or kill it.

This is a creation myth, not science. No operational system, no game theoretic model, no mathematical proof to support these claims has ever been developed. The AGI construct is not a plausible catastrophic risk scenario, but a dark God vision ginned up by a sect of computer scientists who are heavily overrepresented in the field of machine learning and AI.

## 5. AI Governance and the AGI Myth

The doomer AGI vision betrays its advocates' poor understanding of social institutions and of the relationship between technology and society. This is not to say that advances in cyber-physical systems, computer decision making, and generative AI don't pose new and sometimes challenging societal and policy issues. Doubtless, they do. Cybernetic systems will go awry here and there, just as electromechanical systems or civil engineering occasionally do. Humans can misuse AI (or any technology) in many ways. The specter of an autonomous, all-powerful AGI, however, distracts us from the real problems of digital governance. If our threat model is unrealistic, our policy responses are certain to be wrong.

Killing off the AGI Myth – pointing out that the AI doomer's God is dead – is necessary if we are to govern digital technology properly. Most of the talk about AI safety assumes that it is the computer systems themselves that pose the risk. This misperception is based on the AGI myth, which tells us that machine intelligence could become a competing life form that might appear instantly and take over the world.

If this is the way one conceives of the policy problem, only two strategies can be proposed: 1) stop all development, instantly, so that we are sure we will never travel down that path; or 2) micro-regulate the hell out of AI models and algorithms - which turns out to require regulating all computing, networks, data and software (the entire digital ecosystem).

The first option, a "pause AI" agreement, is obviously not feasible. It would require every government, every firm in the computer industry, and everyone in the world with access to ICTs, to agree to abide by the ban. Even defining what activities are and are not banned would be a complex undertaking. Getting all of industry to follow this ban would require overwhelming support from the business community combined with very strict, comprehensive enforcement from all governments. But none of businesses or governments really want to stop, and even if they did the U.S., British, Chinese, European, Russian and Indian governments will not trust each other to stop unilaterally, and hence there would be no enforcement. Given the (alleged) military and business advantages of being the leader in AI technology, each state would have a strong incentive (payoffs for our game theorists) to defect from a pause agreement. There would be no reliable way to detect noncompliance, and if noncompliance was detected, there would be no way for one government or group of governments to enforce the agreement upon another. Unless there is a Hobbesian Leviathan able to wave his scepter over the entire world and command obedience, “pause AI” doesn’t work.

That leaves us with the second option: regulating AI technology. If we think we are forestalling or avoiding human extinction, and we think the autonomous development of the technology itself creates catastrophic risk, then the focus of governance is on precautionary control of the production of AI models and applications. An elite priesthood must be empowered to peer into the inner workings of the algorithms, control the sources of data,[12] control the sources of compute,[13] and engineer the proper values into it. Sastry, Heim, et al, (2024) argue openly that because it is the most centralized aspect of the digital value chain, computing power should be controlled, just as cryptography was controlled as if it were a nuclear weapon in its early days.

Two leading AGI doomers articulate this policy vision well: "in order to avoid catastrophic risks or worse, it is not enough to ensure that only some AGIs are safe. Proposals which seek to solve the issue of catastrophic AGI risk need to also provide some mechanism for ensuring that most (or perhaps even 'nearly all') AGIs are either created safe or prevented from doing considerable harm." (Sotala & Yampolskiy, 2015) Here we have a rationale for comprehensive, globally uniform regulation of the digital ecosystem. If you do not have total control over the production of information technology, then all is lost, because it will be possible for unsafe AGIs to develop and proliferate somewhere.

This is the path we are on now, and while (unlike pause AI) we can proceed with attempts to implement it, it too, is not workable.

Let’s begin with a simple fact: AI is already regulated. Digital technology is heavily institutionalized. Different aspects of it are governed by a large number of interconnected but decentralized social systems: communities of scientific researchers in universities and businesses; telecommunication infrastructure providers, data centers, data sources, privacy law, the price system, venture capital markets, stock markets, governmental regulatory agencies, copyright and patent law, military and civilian funding agencies, some global industry consortia and standards bodies. Our markets and our politics and our choices as consumers are already shaping it. It is false to conceive of an AGI as something outside of us, free of all social and material constraints. This view undermines real AI safety by absolving humans of responsibility for how it evolves and how it is used. And that’s what we need to focus on.

AI – in other words, computing – is a product of society, not the other way around. The current governance paradigm is developing specialized laws and regulations on AI, which

[12] Policies and proposals favoring data localization, restricting the international distribution of apps or social media services, cross-border restrictions on data flows, battles over AI and copyright, all are implicated in any attempt to regulate the data component of AI.

[13] Sastry, Heim, et al, (2024) explicitly target compute for regulation because it is the most concentrated capability (confined to a few major semiconductor producers and cloud providers.

misses the target. AI is not a “thing” that can be isolated and regulated. It is a combination of computing, software, data and networks, and as such touches on all aspects of information and communications and therefore all aspects of society. The current idea, embedded in the U.S. Executive Order, is that if you control the design of AI and regulate its production in the name of “safety,” the risks can be eliminated and nothing wrong can be done with it. Many believe you can put governance “in” the technology. This approach has got it all backwards. It is the users and the uses, not the design, that you need to pay attention to. Designs will be dictated by the incentives of users and uses. If there is a demand for AI, if significant segments of society find specific applications of it useful, those applications will be produced, whether it conforms to some governments’ or some doomers idea of safety or not. Computers, chips, data and networks are ubiquitous and all kinds of applications of them to AI problems already exist.

Most applications of AI are relatively benign and limited in scope. Chatbots are having a substantial impact on the way we write messages in certain genres, and on the way we educate and entertain. It is interesting but it is not life-threatening. It is just another computer application. New applications of computer intelligence have been coming along steadily for 40-50 years. Policy makers concerned with governance of technology need to focus on the context in which the technology is applied and the risks and benefits specific to those situations. Medical applications will have entirely different concerns than law enforcement applications. Before placing a centralized AI application in control of the entire electrical power grid, for example, we need to ask and answer questions about how it will affect the resilience of the grid, about who will assume liability for malfunctions, about its security against external attacks, and about what gains in efficiency will be achieved. We will not answer these questions by controlling the distribution of compute power or by licensing AI models before they are put into use. Those measures are based on the false premise that the machine might come to life and take over the world.

We also need to be more realistic about where major societal risks come from. If Artificial Intelligence ever becomes a systemic threat to human life, it will not be because a new race of cyber-terminators emerges spontaneously from a computer science lab. It most likely will come from an arms race prompted by military conflicts among nation-states, in which one state actor’s pursuit of supremacy makes it indifferent to damage to the citizens of the other state actor. Some threats will certainly arise from poor implementations and poor designs, and some from criminal exploitation of its capabilities. But whatever the risks are, in all cases humans, not AI, generate the threat. None of these threats pose a risk of human extinction, unless humans themselves are pursuing the goal of human extinction. War is the only scenario in which that is a plausible outcome.

Institutionalized competition for political and military power is the biggest threat. Each state will utilize and develop artificial intelligence in a way that best serves its own interests, and since states are in the business of maintaining a monopoly on the use of force, out-of-control competition over AI capabilities among human governmental armed forces is what we need to look out for. Ironically, the AI doom scenarios reinforce this risk by suggesting that

AGI development will yield omnipotent powers. This only encourages governments to see it as a prospective weapon and to seek exclusive control of it.

Whether or not political-military competition among states leads to societal harm depends not on machine evolution, but on social evolution – i.e., on how we structure our interactions. More progress is likely to come from properly governing global trade, and from avoiding international conflict among nation states, than from micro-interventions in the design and distribution of computing technology.

By decoupling their digital ecosystems, the United States and China are contributing to the kind of competition that will push digital technology into threatening paths. We are turning our backs on the economic and social benefits of a globally interconnected, interoperable digital ecosystem. We have lost sight of its contribution to cooperation, peace and prosperity. In a digital ecosystem fragmented by geopolitics, ICT becomes not a means of global communication, cooperation and commerce, but a weapon. All information systems become weapons, in fact. Social media, data, applications, networks, semiconductors, even batteries and EVs all are brought into the service of the national security state. The evolution of hostile and destructive applications of AI and cyber-physical systems is more likely to come from military competition for "supremacy" in AI than from some self-generating, self-replicating superintelligence. The long-term danger is not AI technology per se, but the way its development and supply chain are incorporated into geopolitical rivalries and weapons systems controlled by states.

## References

Brockman, J. (ed). (2019). *Possible Minds: 25 Ways of Looking at AI*. Penguin.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... & Ng, A. (2012). Large scale distributed deep networks. *Advances in neural information processing systems*, *25*.

Delipetrev, B., Tsinaraki, C., and Kostic, U. (2020) *Historical Evolution of Artificial Intelligence*. Technical Report. Publications Office of the European Union.

Foss, N. J. (1996). Firms, incomplete contracts, and organizational learning. *Human Systems Management*, *15*(1), 17-26.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on information systems (TOIS), 14(3), 330-347.

Hadfield-Menell, D., Dragan, A., Abbeel, P. Russell. S. 2017. The Off-Switch Game. The AAAI-17 Workshop on AI, Ethics, and Society. https://cdn.aaai.org/ocs/ws/ws0354/15156-68335-1-PB.pdf

Jones, C. (2023) "The A.I. Dilemma: Growth versus Existential Risk." Stanford GSB and NBER. https://web.stanford.edu/~chadj/existentialrisk.pdf

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*(1), 90-96.

Kurzweil, R. (1999). Spiritual machines. Research & Development, 41(7), 14-18

Leslie, D. (2020). Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023*.

McMillan, R. & Seetharaman, D. (2023). How a Fervent Belief Split Silicon Valley—and Fueled the Blowup at OpenAI." Wall Street Journal, Nov. 22, 2023 2:25 pm ET.

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Omohundro, S. M. “The nature of self-improving artificial intelligence.” http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/, October 2007.

Omohundro, S. M. (2008). The basic AI drives. pp. 47-55 in Yampolskiy, R (ed.). (2018). Artificial intelligence safety and security. Chapman and Hall/CRC.

Orike, S., & Ene, D. S. (2023). Meta-Learning: Unleashing the Power of Self-Improving Artificial Intelligent (AI) Systems. *Journal of Advances in Computational Intelligence Theory*, *5*(3), 12-27.

Pinker, S. (2019). "Tech Prophecy and the Underappreciated Causal Power of Ideas," Chapter 10 in Brockman, 2019.

Reddy, P. P. (2020). Artificial Superintelligence: A Model for Self-Improving / Self-Modifying Programs, EasyChair preprint,

Shulman, Carl. 2010. Omohundro's "Basic AI Drives" and Catastrophic Risks. MACHINE INTELLIGENCE RESEARCH INSTITUTE (MIRI).

Sheehan, M. (July 10, 2023) China's AI Regulations and How They Get Made. Working Paper, Carnegie Endowment for International Peace https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117

Spiller, P. T. (2009). An institutional theory of public contracts: Regulatory implications. In Regulation, deregulation, reregulation. Edward Elgar Publishing.

von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. (Edited and completed by Arthur Burks). Urbana and London: University of Illinois Press.

Wängberg, T., Böörs, M., Catt, E., Everitt, T., & Hutter, M. (2017). A game-theoretic analysis of the off-switch game. In *Artificial General Intelligence: 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings 10* (pp. 167-177). Springer International Publishing.

Zhang, A. H. (2024). The Promise and Perils of China's Regulation of Artificial Intelligence. *Available at SSRN*.