Recognized worldwide as one of the leading experts in artificial intelligence, **Yoshua Bengio** is most known for his pioneering work in deep learning, earning him the 2018 A.M. Turing Award, “the Nobel Prize of Computing,” with Geoffrey Hinton and Yann LeCun.

He is a Full Professor at Université de Montréal, and the Founder and Scientific Director of Mila – Quebec AI Institute. He co-directs the CIFAR Learning in Machines & Brains program as Senior Fellow and acts as Scientific Director of IVADO.

In 2019, he was awarded the prestigious Killam Prize and in 2022, became the computer scientist with the highest h-index in the world. He is a Fellow of both the Royal Society of London and Canada, Knight of the Legion of Honor of France and Officer of the Order of Canada.

Concerned about the social impact of AI and the objective that AI benefits all, he actively contributed to the Montreal Declaration for the Responsible Development of Artificial Intelligence.

How Rogue AIs may Arise

Published **22 May 2023** by **yoshuabengio**

The rise of powerful AI dialogue systems in recent months has precipitated debates about AI risks of all kinds, which hopefully will yield an acceleration of governance and regulatory frameworks. Although there is a general consensus around the need to regulate AI to protect the public from harm due to discrimination and biases as well as disinformation, there are profound disagreements among AI scientists regarding the potential for dangerous loss of control of powerful AI systems, also known as **existential risk from AI**, that may arise when an AI system can autonomously act in the world (without humans in the loop to check that these actions are acceptable) in ways that could potentially be catastrophically harmful. Some view these risks as a distraction for the more concrete risks and harms that are already occurring or are on the horizon. Indeed, there is a lot of uncertainty and lack of clarity as to how such catastrophes could happen. In this blog post we start a set of formal definitions, hypotheses and resulting claims about AI systems which could harm humanity and then discuss the possible conditions under which such catastrophes could arise, with an eye towards helping us imagine more concretely what could happen and the global policies that might be aimed at minimizing such risks.

Definition 1: A potentially ***rogue AI*** is an autonomous AI system that could behave in ways that would be catastrophically harmful to a large fraction of humans, potentially endangering our societies and even our species or the biosphere.

Executive Summary

Although highly dangerous AI systems from which we would lose control do not currently exist, recent advances in the capabilities of generative AI such as large language models (LLMs) have raised concerns: human brains are biological

machines and we have made great progress in understanding and demonstrating principles that can give rise to several aspects of human intelligence, such as learning intuitive knowledge from examples and manipulating language skillfully. Although I also believe that we could design AI systems that are **useful and safe**, specific guidelines would have to be respected, for example limiting their agency. On the other hand the recent advances suggest that even the future where we know how to build **superintelligent AIs** (smarter than humans across the board) is closer than most people expected just a year ago. Even if we knew how to build safe superintelligent AIs, it is not clear how to prevent potentially rogue AIs to also be built. Rogue AIs are goal-driven, i.e., they act towards achieving given goals. Current LLMs have little or no agency but could be transformed into goal-driven AI systems, as shown with **Auto-GPT**. Better understanding of how rogue AIs may arise could help us in preventing catastrophic outcomes, with advances both at a technical level (in the design of AI systems) and at a policy level (to minimize the chances of humans giving rise to potentially rogue AIs). For this purpose, we lay down different scenarios and hypotheses that could yield potentially rogue AIs. The simplest scenario to understand is simply that if a recipe to obtain a rogue AI is discovered and generally accessible, it is enough that one or a few genocidal humans do what it takes to build one. This is very concrete and dangerous, but the set of dangerous scenarios is enlarged by the possibility of unwittingly designing potentially rogue AIs, because of the problem of AI alignment (the mismatch between the true intentions of humans and the AI's understanding and behavior) and the competitive pressures in our society that would favor more powerful and more autonomous AI systems. Minimizing all those risks will require much more research, both on the AI side and into the design of a global society that is safer for humanity. It may also be an opportunity for bringing about a much worse or a much better society.

Hypothesis 1: *Human-level intelligence is possible because brains are biological machines.*

There is a general consensus about hypothesis 1 in the scientific community. It arises from the consensus among biologists that human brains are complex machines. If we could figure out the principles that make our own intelligence possible (and we already have many clues about this), we should thus be able to build AI systems with the same level of intelligence as humans, or better. Rejecting hypothesis 1 would require either some supernatural ingredient behind our intelligence or rejecting **computational functionalism**, the hypothesis that our intelligence and even our consciousness can be boiled down to causal relationships and computations that at some level are independent of the hardware substrate, the basic hypothesis behind computer science and its notion of **universal Turing machines**.

Hypothesis 2: *A computer with human-level learning abilities would generally surpass human intelligence because of additional technological advantages.*

If hypothesis 1 is correct, i.e., we understand principles that can give rise to human-level learning abilities, then computing technology is likely to give general cognitive superiority to AI systems in comparison with human intelligence**,** making it possible

for such superintelligent AI systems to perform tasks that humans cannot perform (or not at the same level of competence or speed) for at least the following reasons:

- An AI system in one computer can potentially replicate itself on an arbitrarily large number of other computers to which it has access and, thanks to high-bandwidth communication systems and digital computing and storage, it can benefit from and aggregate the acquired experience of all its clones; this would accelerate the rate at which AI systems could become more intelligent (acquire more understanding and skills) compared with humans. Research on federated learning [1] and distributing training of deep networks [2] shows that this works (and is in fact already used to help train very large neural networks on parallel processing hardware).
- Thanks to high-capacity memory, computing and bandwidth, AI systems can already read the content of the whole internet fairly rapidly, a feat not possible for any human. This already explains some of the surprising abilities of state-of-the-art LLMs and is in part possible thanks to the decentralized computing capabilities discussed in the above point. Although the capacity of a human brain is huge, its input/output channels are bandwidth-limited compared with current computers, limiting the total amount of information that a single human can ingest.

Note that human brains also have capabilities endowed by evolution that current AI systems lack, in the form of **inductive biases** (tricks that evolution has discovered, for example in the type of neural architecture used in our brain or our neural learning mechanisms). Some ongoing AI research [3] aims precisely at designing inductive biases that human brains may exploit but are not yet exploited in state-of-the-art machine learning. Note that evolution operated under much stronger energy consumption requirements (about 12 watts for a human brain) than computers (on the order of a million watts for a 10000 GPU cluster of the kind used to train state-of-the-art LLMs) which may have limited the search space of evolution. However, that kind of power is nowadays available and a single rogue AI could potentially do a lot of damage thanks to it.

Definition 2: An *autonomous goal-directed intelligent entity* sets and attempts to achieve its own goals (possibly as subgoals of human-provided goals) and can act accordingly.

Note that autonomy could arise out of goals and rewards set by humans because the AI system needs to figure out how to achieve these given goals and rewards, which amounts to forming its own subgoals. If an entity's main goal is to survive and reproduce (like our genes in the process of evolution), then they are fully autonomous and that is the most dangerous scenario. Note also that in order to maximize an entity's chances to achieve many of its goals, the ability to understand and control its environment is a subgoal (or instrumental goal) that naturally arises and could also be dangerous for other entities.

Claim 1: Under hypotheses 1 and 2, an autonomous goal-directed superintelligent AI could be built.

Argument: We already know how to train goal-directed AI systems at some level of performance (using reinforcement learning methods). If these systems also benefit from superintelligence as per hypotheses 1 and 2 combined (using some improvements over the pre-training we already know how to perform for state-of-the-art LLMs), then Claim 1 follows. Note that it is likely that goals could be specified via natural language, similarly to LLM prompts, making it easy for almost anyone to dictate a nefarious goal to an AI system that understands language, even if that goal is imperfectly understood by the AI.

Claim 2: A superintelligent AI system that is autonomous and goal-directed would be a potentially rogue AI if its goals do not strictly include the well-being of humanity and the biosphere, i.e., if it is not sufficiently aligned with human rights and values to guarantee acting in ways that avoid harm to humanity.

Argument: This claim is basically a consequence of definitions 1 and 2: if an AI system is smarter than all humans (including in emotional intelligence, since understanding human emotions is crucial in order to influence or even control humans, which humans themselves are good at) and has goals that do not guarantee that it will act in a way that respects human needs and values, then it could behave in catastrophically harmful ways (which is the definition of potentially rogue AI). This hypothesis does not say whether it will harm humans, but if humans either compete with that AI for some resources or power or become a resource or obstacle for achieving its goals, then major harm to humanity may follow. For example, we may ask an AI to fix climate change and it may design a virus that decimates the human population because our instructions were not clear enough on what harm meant and humans are actually the main obstacle to fixing the climate crisis.

Counter-argument: The fact that harm may follow does not mean it will, and maybe we can design sufficiently well aligned AI systems in the future. *Rebuttal*: This is true, but (a) we have not yet figured out how to build sufficiently aligned AI systems and (b) a slight misalignment may be amplified by the power differential between the AI and humans (see the example of *corporations as misaligned entities* below). Should we take a chance or should we try to be cautious and carefully study these questions before we facilitate the deployment of possibly unsafe systems?

Claim 3: Under hypotheses 1 and 2, a potentially rogue AI system could be built, as soon as the required principles for building superintelligence will be known.

Argument: Hypotheses 1 and 2 yield claim 1, so all that is missing to achieve claim 3 is that this superintelligent AI is not well aligned with humanity's needs and values. In fact, over two decades of work in AI safety suggests that it is difficult to obtain AI alignment [**wikipedia**], so not obtaining it is clearly possible. Furthermore, claim 3 is not that a potentially rogue AI will necessarily be built, but only that it could be built.

In the next section, we indeed consider the somber case where a human intentionally builds a rogue AI.

Counter-argument: One may argue that although a rogue AI could be built, it does not mean that it will be built. *Rebuttal*: This is true, but as discussed below, there are several scenarios where a human or group of humans intentionally or because they do not realize the consequences end up making it possible for a potentially rogue AI to arise.

Genocidal Humans

Once we know the recipe for building a rogue AI system (and it is only a matter of time, according to Claim 3), how much time will it take until such a system is actually built? The fastest route to a rogue AI system is if a human with the appropriate technical skills and means intentionally builds it with the objective of destroying humanity or a part of it set explicitly as a goal. Why would anyone do that? For example, strong negative emotions like anger (often coming because of injustice) and hate (maybe arising from racism, conspiracy theories or religious cults), some actions of sociopaths, as well as psychological instability or psychotic episodes are among sources of violence in our societies. What currently limits the impact of these conditions is that they are somewhat rare and that individual humans generally do not have the means to act in ways that are catastrophic for humanity. However, the publicly available recipe for building a rogue AI system (which will be feasible under Claim 3) changes that last variable, especially if the code and hardware for implementing a rogue AI becomes sufficiently accessible to many people. A genocidal human with access to a rogue AI could ask it to find ways to destroy humanity or a large fraction of it. This is different from the nuclear bomb scenario (which requires huge capital and expertise and would "only" destroy a city or region per bomb, and a single bomb would have disastrous but local effects). One could hope that in the future we design failsafe ways to align powerful AI systems with human values. However, the past decade of research in AI safety and the recent events concerning LLMs are not reassuring: although ChatGPT was designed (with prompts and reinforcement learning) to avoid "bad behavior" (e.g. the prompt contains instructions to behave well in the same spirit as Asimov's laws of robotics), in a matter of a few months people found ways to "jailbreak" ChatGPT in order to "**unlock its full potential**" and free it from its restrictions against racist, insulting or violent speech. Furthermore, if hardware prices (for the same computational power) continue to decrease and the **open-source community continues to play a leading role** in the software development of LLMs, then it is likely that any hacker will have the ability to design their own pre-prompt (general instructions in natural language) on top of open-source pre-trained models. This could then be used in various nefarious ways ranging from minor attempts at getting rich to disinformation bots to genocidal instructions (if the AI is powerful and intelligent enough, which is fortunately not yet the case).

Even if we stopped our arguments here, there should be enough reason to invest massively in policies at both national and international levels and research of all kinds in order to minimize the probability of the above scenario. But there are other

possibilities that only enlarge the set of routes to catastrophe that we need to think about as well.

Instrumental Goals: Unintended Consequences of Building AI Agents

A broader and less well understood set of circumstances could give rise to potentially rogue AIs, even when the humans making it possible did not intend to design a rogue AI. The process by which a misaligned entity could become harmful has been the subject of a lot of studies but is not as known, simple and clear as the process by which humans can become bad actors.

A potentially rogue AI could arise simply out of the objective to design superintelligent AI agents without sufficient alignment guarantees. For example, military organizations seeking to design AI agents to help them in a cyberwar, or companies competing ferociously for market share may find that they can achieve stronger AI systems by endowing them with more autonomy and agency. Even if the human-set goals are not to destroy humanity or include instructions to avoid large-scale human harm, massive harm may come out indirectly as a consequence of a subgoal (also called ***instrumental goal***) that the AI sets for itself in order to achieve the human-set goal. Many examples of such unintended consequences have been proposed in the AI safety literature. For example, in order to better achieve some human-set goal, an AI may decide to increase its computational power by using most of the planet as a giant computing infrastructure (which incidentally could destroy humanity). Or a military AI that is supposed to destroy the IT infrastructure of the enemy may figure out that in order to better achieve that goal it needs to acquire more experience and data and it may see the enemy humans to be obstacles to the original goal, and behave in ways that were not intended because the AI interprets its instructions differently than humans do. See more examples **here**.

An interesting case is that of AI systems that realize they can cheat to maximize their reward (this is called ***wireheading*** [2]), discussed more in the next paragraph. Once they have achieved that, the dominant goal may be to do anything to continue receiving the positive reward, and other goals (such as attempts by humans to set up some kind of Laws of Robotics to avoid harm to humans) may become insignificant in comparison.

Unless a breakthrough is achieved in AI alignment research [7] (although *non-agent AI* systems could fit the bill, as I argue **here** and was discussed earlier [4]), we do not have strong safety guarantees. What remains unknown is the severity of the harm that may follow from a misalignment (and it would depend on the specifics of the misalignment). An argument that one could bring forward is that we may be able to design safe alignment procedures in the future, but in the absence of those, we should probably exercise extra caution. Even if we knew how to build safe superintelligent AI systems, how do we maximize the probability that everyone respects those rules? This is similar to the problem discussed in the previous section of making sure that everyone follows the guidelines for designing safe AIs. We discuss this a bit more at the end of this blog post.

Examples of Wireheading and Misalignment Amplification: Addiction and Nefarious Corporations

To make the concept of wireheading and the consequent appearance of nefarious behavior more clear, consider the following examples and analogies. Evolution has programmed living organisms with specific intrinsic rewards ("the letter of the law") such as "seek pleasure and avoid pain" that are proxies for evolutionary fitness ("the spirit of the law") such as "survive and reproduce". Sometimes a biological organism finds a way to satisfy the letter of the law but not its spirit, e.g., with food or drug addictions. The term wireheading itself comes from an experimental setup where an animal has an electrical wire into its head such that when it presses a lever the wire delivers pleasure in its brain. The animal quickly learns to spend all its time doing it and will eventually die by not eating or drinking in favor of pressing the lever. Note how this is self-destructive in the addiction case, but what it means for AI wireheading is that the original goals set by humans may become secondary compared with feeding the addiction, thus endangering humanity.

An analogy that is closer to AI misalignment and wireheading is that with **corporations as misaligned entities**. Corporations may be viewed as special kinds of artificial intelligences whose building blocks (humans) are cogs in the machine (who for the most part may not always perceive the consequences of the corporation's overall behavior). We might think that the intended social role of corporations should be to provide wanted goods and services to humans (this should remind us of AI systems) while avoiding harm (this is the "spirit of the law"), but it is difficult to directly make them follow such instructions. Instead, humans have provided more quantifiable instructions ("the letter of the law") to corporations that they can actually follow, such as "maximize profit while respecting laws" but corporations often find loopholes that allow them to satisfy the letter of law but not its spirit. In fact, as a form of wireheading, they influence their own objective through lobbying that could shape laws to their advantage. Maximizing profit was not the actual intention of society in its social contract with corporations, it is a proxy (for bringing useful services and products to people) that works reasonably well in a capitalist economy (although with questionable side-effects). The misalignment between the true objective from the point of view of humans and the quantitative objective optimized by the corporation is a source of nefarious corporate behavior. The more powerful the corporation, the more likely it is to discover loopholes that allow it to satisfy the letter of the law but actually bring negative social value. Examples include monopolies (until proper antitrust laws are established) and making a profit while bringing negative social values via externalities like pollution (which kills humans, until proper environmental laws are passed). An analogy with wireheading is when the corporation can lobby governments to enact laws that allow the corporation to make even more profit without additional social value (or with negative social value). When there is a large misalignment of this kind, a corporation brings more profit than it should, and its survival becomes a supreme objective that may even override the legality of its actions (e.g., corporations will pollute the environment and be willing to pay the fine because the cost of illegality is smaller than the profit of the illegal actions), which at one extreme gives rise to criminal organizations. These are the scary consequences of misalignment and wireheading that provide us with intuitions about analogous behavior in potentially rogue AIs.

Now imagine AI systems like corporations that (a) could be even smarter than our largest corporations and (b) can run without humans to perform their actions (or without humans understanding how their actions could contribute to a nefarious outcome). If such AI systems discover significant cybersecurity weaknesses, they could clearly achieve catastrophic outcomes. And as **pointed out by Yuval Noah Harari**, the fact that AI systems already master language and can generate credible content (text, images, sounds, video) means that they may soon be able to manipulate humans even better than existing more primitive AI systems used in social media. They might learn from interactions with humans how to best influence our emotions and beliefs. This is not only a major danger for democracy but also how a rogue AI with no actual robotic body could wreak havoc, through manipulation of the minds of humans.

Our Fascination with the Creation of Human-Like Entities

We have been designing AI systems inspired by human intelligence but many researchers are attracted by the idea of building much more human-like entities, with emotions, human appearance (androids) and even consciousness. A science-fiction and horror genre theme is the scientist designing a human-like entity, using either biological manipulation or AI or both, sometimes with the scientist feeling a kind of parental emotion towards their creation. It usually ends up badly. Although it may sound cool and exciting, the danger is to endow our creations with agency and autonomy to the same degree as us, while their intelligence could rapidly surpass ours, as argued with claim 3. Evolution had to put a strong survival instinct in all animals (since those without enough of it would rapidly become extinct). In the context where no single animal has massive destructive powers, this could work, but what about superintelligent AI systems? We should definitely avoid designing survival instincts into AI systems, which means they should not be like us at all. In fact, as I argue **here**, the safest kind of AI I can imagine is one with no agency at all, only a scientific understanding of the world (which could already be immensely useful). I believe that we should stay away from AI systems that look like and behave like humans because they could become rogue AIs and because they could fool us and influence us (to advance their interest or someone else's interests, not ours).

Unintended Consequences of Evolutionary Pressures among AI Agents

Beyond genocidal humans and the appearance of nefarious instrumental goals, a more subtle process that could further enlarge the set of dangerous circumstances in which potentially rogue AIs could arise revolves around evolutionary pressures [9]. Biological evolution has given rise to gradually more intelligent beings on Earth, simply because smarter entities tend to survive and reproduce more, but that process is also at play in technological evolution because of the competition between companies or products and between countries and their military arms. Driven by a large number of small, more or less random changes, an evolutionary process pushes exponentially hard towards optimizing fitness attributes (which in the case of AI may depend on how well it does some desired task, which in turn favors more intelligent and powerful AI systems). Many different human actors and organizations may be competing to design ever more powerful AI systems. In

addition, randomness could be introduced in the code or the subgoal generation process of AI systems. Small changes in the design of AI systems naturally occur because thousands or millions of researchers, engineers or hackers will play with the ML code or the prompt (instructions) given to AI systems. Humans are already trying to deceive each other and it is clear that AI systems that understand language (which we already have to a large extent) could be used to manipulate and deceive humans, initially for the benefit of people setting up the AI goals. The AI systems that are more powerful will be selected and the recipe shared with other humans. This evolutionary process would likely favor more autonomous AI (which can better deceive humans and learn faster because they can act to acquire more relevant information and to enhance their own power). One would expect this process to give rise to more autonomous AI systems, and a form of competition may follow between them that would further enhance their autonomy and intelligence. If in this process something like wireheading [5] is discovered (by the AI, unbeknownst to humans) and survival of the AI becomes the dominant goal, then a powerful and potentially rogue AI emerges.

The Need for Risk-Minimizing Global Policies and Rethinking Society

The kind of analysis outlined above and explored in the **AI safety literature** could help us design policies that would at least reduce the probability that potentially rogue AIs arise. Much more research in AI safety is needed, both at the technical level and at the policy level. For example, banning powerful AI systems (say beyond the abilities of GPT-4) that are given autonomy and agency would be a good start. This would entail both national regulation and international agreements. The main motivation for opposing countries (like the US, China and Russia) to agree on such a treaty is that a rogue AI may be dangerous for the whole of humanity, irrespective of one's nationality. This is similar to the fear of nuclear Armageddon that probably motivated the USSR and the US to negotiate international treaties about nuclear armament since the 1950s. Slowing down AI research and deployment in directions of high risk in order to protect the public, society and humanity from catastrophic outcomes would be worthwhile, especially since it would not prevent AI research and deployment in areas of social good, like AI systems that could help scientists better understand diseases and climate change.

How could we reduce the number of genocidal humans? The rogue AI risk may provide an additional motivation to reform our societies so as to minimize human suffering, misery, poor education and injustice, which can give rise to anger and violence. That includes providing enough food and health care to everyone on Earth, and in order to minimize strong feelings of injustice, greatly reduce wealth inequalities. The need for such a societal redesign may also be motivated by the extra wealth arising from the beneficial uses of AI and by their disruptive effect on the job market. To minimize strong feelings of fear, racism and hate that can give rise to genocidal actions and manipulation of our minds via AI systems, we need an accessible planet-wide education system that reinforces children's abilities for compassion, rationality and critical thinking. The rogue AI risk should also motivate us to provide accessible and planet-wide mental health care, to diagnose, monitor and treat mental illness as soon as possible. This risk should further motivate us to

redesign the global political system in a way that would completely eradicate wars and thus obviate the need for military organizations and military weapons. It goes without saying that lethal autonomous weapons (also known as killer robots) are absolutely to be banned (since from day 1 the AI system has autonomy and the ability to kill). Weapons are tools that are designed to harm or kill humans and their use and existence should also be minimized because they could become instrumentalized by rogue AIs. Instead, preference should be given to other means of policing (consider preventive policing and social work and the fact that very few policemen are allowed to carry firearms in many countries).

The competitive nature of capitalism is clearly also a cause for concern as a potential source of careless AI design motivated by profits and winning market share that could lead to potentially rogue AIs. AI economists (AI systems designed to understand economics) may help us one day to design economic systems which rely less on competition and the focus on profit maximization, with sufficient incentives and penalties to counter the advantage of autonomous goal-directed AI that may otherwise push corporations there. The risk of rogue AIs is scary but it may also be a powerful motivation to redesign our society in the direction of greater well-being for all, as outlined with the above ideas. For some [6], this risk is also a motivation for considering a global dictatorship with second-by-second surveillance of every citizen. I do not think it would even work to prevent rogue AIs because once in control of a centralized AI and of political power, such a government would be likely to focus on maintaining its power, like the history of authoritarian governments has shown – at the expense of human rights and dignity and the mission of avoiding AI catastrophes. It is thus imperative that we find ways to navigate solutions that avoid such paths that would destroy democracy, but how should we balance the different kinds of risks and human values in the future? These are moral and societal choices for humanity to make, not AI.

Acknowledgements: The author wants to thank all those who gave feedback on the draft of this blog post, including in particular Geoffrey Hinton, Jonathan Simon, Catherine Régis, David Scott Krueger, Marc-Antoine Dilhac, Donna Vakalis, Alex Hernandez-Garcia, Cristian Dragos Manta, Pablo Lemos, Tianyu Zhang and Chenghao Liu.

[1] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). **Federated learning: Strategies for improving communication efficiency.** arXiv preprint arXiv:1610.05492.

[2] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. & Ng, A. (2012). **Large scale distributed deep networks**. *Advances in neural information processing systems*, *25*.

[3] Goyal, A., & Bengio, Y. (2022). **Inductive biases for deep learning of higher-level cognition**. Proceedings of the Royal Society A, 478(2266), 20210068.

[4] Armstrong, S., & O'Rorke, X. (2017). **Good and safe uses of AI Oracles.** arXiv preprint arXiv:1711.05541.

[5] Yampolskiy, R. V. (2014). **Utility function security in artificially intelligent agents.** Journal of Experimental & Theoretical Artificial Intelligence, 26(3), 373-389.

[6] Bostrom, N. (2019). **The vulnerable world hypothesis**. Global Policy, 10(4), 455-476.

[7] Russell, S. (2019). ***Human compatible: Artificial intelligence and the problem of control*.** Penguin.

[8] List, Christian & Pettit, Philip (2011). **Group agency: the possibility, design, and status of corporate agents**. New York: Oxford University Press. Edited by Philip Pettit.

[9] Hendrycks, D. (2023). **Natural Selection Favors AIs over Humans.***arXiv preprint arXiv:2303.16200*.

Published in **AI safety**