**Yoshua Bengio**

# Reasoning through arguments against taking AI safety seriously

Published **9 July 2024** by **yoshuabengio**

About a year ago, a few months after I **publicly took a stand** with many other peers to warn the public of the dangers related to the unprecedented capabilities of powerful AI systems, I posted a blog post entitled **FAQ on Catastrophic AI Risks** as a follow-up to my earlier **one about rogue AIs**, where I started discussing why AI safety should be taken seriously. In the meantime, I participated in numerous debates, including many with my friend Yann LeCun, whose views on some of these issues are very different from mine. I also learned a lot more about AI safety, how different groups of people think about this question as well as the diversity of views about regulation and the efforts of powerful lobbies against it. The issue is so hotly debated because the stakes are major: According to some estimates, **quadrillions of dollars** of net present value are up for grabs, not to mention political power great enough to significantly disrupt the current world order. I published a **paper on multilateral governance of AGI labs** and I spent a lot of time thinking about catastrophic AI risks and their mitigation, both on the technical side and the governance and political side. In the last seven months, I have been chairing (and continue to chair) the **International Scientific Report on the Safety of Advanced AI** ("the report", below), involving a panel of 30 countries plus the EU and UN and over 70 international experts to synthesize the state of the science in AI safety, illustrating the broad diversity of views about AI risks and trends. Today, after an intense year of wrestling with these critical issues, I would like to revisit arguments made about the potential for catastrophic risks associated with AI systems anticipated in the future, and share my latest thinking.

There are many risks regarding the race by several private companies and other entities towards human-level AI (a.k.a. AGI) and beyond (a.k.a. ASI for Artificial Super-Intellingence). Please see "**the report**" for a broad coverage of risks ranging from current human rights issues to threats on privacy, democracy, copyright, concerns about concentration of economic and political power, and, of course,

dangerous misuse. Although experts may disagree on the probability of various outcomes, we can generally agree that some major risks, like the extinction of humanity for example, would be so catastrophic if they happened that they require special attention, if only to make sure that their probability is infinitesimal. Other risks like severe threats to democracies and human rights also deserve much more attention than they are currently getting.

The most important thing to realize, through all the noise of discussions and debates, is a very simple and indisputable fact: **while we are racing towards AGI or even ASI, nobody currently knows how such an AGI or ASI could be made to behave morally, or at least behave as intended by its developers and not turn against humans**. It may be difficult to imagine, but just picture this scenario for one moment:

*Entities that are smarter than humans and that have their own goals: are we sure they will act towards our well-being?*

Can we collectively take that chance while we are not sure? Some people bring up all kinds of arguments why we should not worry about this (I will develop them below), but they cannot provide a technical methodology for demonstrably and satisfyingly controlling even current advanced general-purpose AI systems, much less guarantees or strong and clear scientific assurances that with such a methodology, an ASI would not turn against humanity. It does not mean that a way to achieve *AI alignment and control* that could scale to ASI could not be discovered, and in fact I argue below that the scientific community and society as a whole should make a massive collective effort to figure it out.

In addition, even if the way to control an ASI was known, **political institutions to make sure that the power of AGI or ASI would not be abused by humans against humans at a catastrophic scale, to destroy democracy or bring about geopolitical and economic chaos or dystopia would still be missing**. *We need to make sure that no single human, no single corporation and no single government can abuse the power of AGI at the expense of the common good*. We need to make sure that corporations do not use AGI to co-opt their governments and governments using it to oppress their people and nations using it to dominate internationally. And at the same time, we need to make sure that we avoid catastrophic accidents of loss of control with AGI systems, anywhere on the planet. All this can be called the *coordination problem,* i.e., the politics of AI. If the coordination problem was solved perfectly, solving the AI alignment and control problem would not be an absolute necessity: we could “just” collectively apply the precautionary principle and avoid

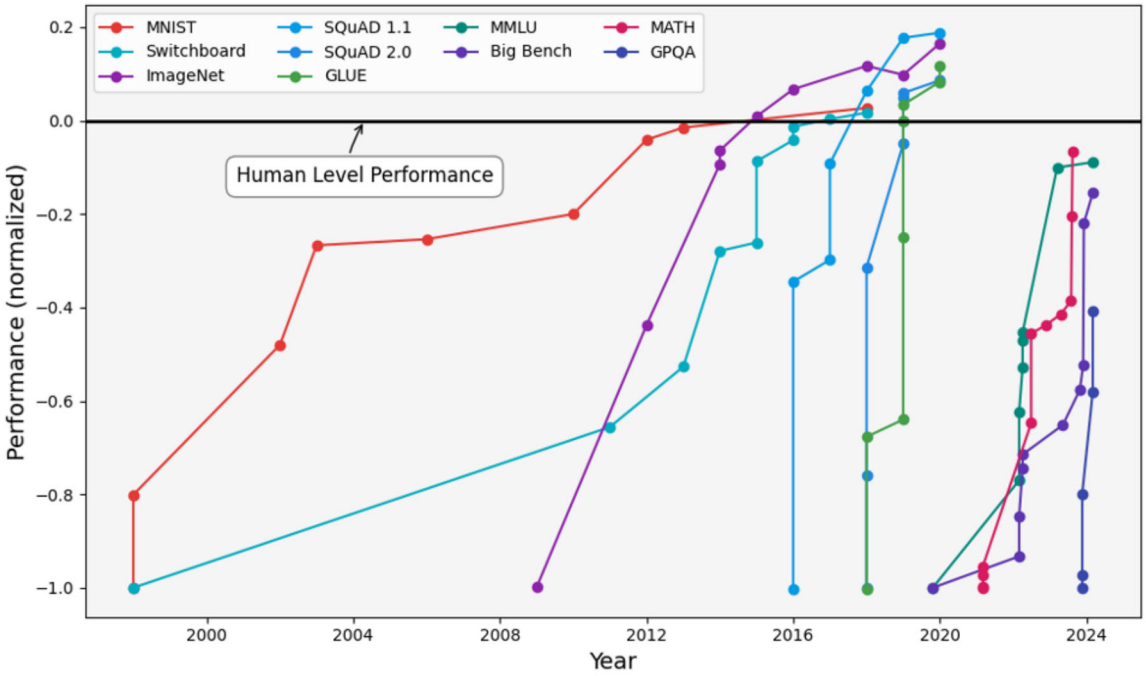
doing experiments anywhere with a non-trivial risk of constructing uncontrolled AGI. But of course, humanity is not a single mind but billions of them, many wills, many countries and corporations each with their objective: The dynamics of all these self-interests and psychological or cultural factors are currently leading us into a dangerous race towards greater AI capabilities without the methodology and institutions to sufficiently mitigate the greatest risks, such as catastrophic misuse and loss of control. And, on a more positive note, *if both the AI control problem and the AI coordination problem are solved*, I buy the argument that there is a good chance that humanity could benefit immensely from the scientific and technological advances that could follow, including in the areas of health, the environment and ensuring better economic prospects for the majority of humans (ideally starting with those who need it most).

As of now, however, we are racing towards a world with entities that are smarter than humans and pursue their own goals – without a reliable method for humans to ensure those goals are compatible with collective human goals. Nonetheless, in my conversations about AI safety I have heard various arguments meant to support a "no worry" conclusion. My general response to most of these arguments is that given the compelling basic case for why the race to AGI could lead to danger – even without certainty, and given the high stakes, we should aim to have very strong evidence before concluding there is nothing to worry about. Often, I find that these arguments fail to meet this bar by a lot. Below, I discuss some of these arguments and why they have not convinced me that we can ignore potential catastrophic risks from AI. Many of the 'no worry' arguments I have heard or read are not actual sound arguments, but intuitions of people who feel certain that there is no danger but offer no convincing chain of reasoning. Without having such arguments to deny the importance of AI safety and when considering our global well-being and the uncertainty about the future, rational decision-making calls for humility, recognizing our epistemic uncertainty and following scientific decision theory, which leads to the precautionary principle. But I feel we are not: Yes, extreme risks from AI are being discussed more now and are not being systematically ridiculed anymore. But we are still not taking them seriously enough. Many people, including decision-makers, are now aware that AI might pose catastrophic and even existential risks. But how vividly do they imagine what this might mean? How willing are they to take unconventional steps to mitigate these risks? I worry that with the current trajectory of public and political engagement with AI risk, we could collectively sleepwalk – even race – into a fog behind which could lie a catastrophe that many knew was possible, but whose prevention wasn't prioritized enough.

## For those who think AGI and ASI are impossible or are centuries in the future

One objection to taking AGI/ASI risk seriously states that we will never (or only in the far future) reach AGI or ASI. Often, this involves statements like "The AIs just predict the next word", "AIs will never be conscious", or "AIs cannot have *true* intelligence". I find most such statements unconvincing because they often conflate two or more concepts and therefore miss the point. For instance, consciousness is not well understood and it is not clear that it is necessary for either AGI or ASI, and it will not necessarily matter for potential existential AGI risk. What will matter most and is more concrete are the capabilities and intentions of ASI systems. If they can kill humans (it's a capability among others that can be learned or deduced from other skills) and have such a goal (and we already have goal-driven AI systems), this could be highly dangerous unless a way to prevent this or countermeasures are found.

I also find statements like "AIs cannot have *true* intelligence" or "The AIs just predict the next word" unconvincing. I agree that if one defines "*true"* intelligence as "the way humans are intelligent", AIs don't have "*true*" intelligence – their way of processing information and reasoning is different from ours. But in a conversation about potential catastrophic AI risks, this is a distraction. What matters for such a conversation is: What can the AI achieve? How good is it at problem-solving? That's how I think of "AGI" and "ASI" – a level of AI capabilities at which an AI is as good as, or better than, a human expert for basically any cognitive task (excluding problems that require physical actions). *How* the AI is capable of this does not change the existence of the risk. And looking at the abilities of AI systems across decades of research, there is a very clear trend of increasing abilities. There is also the current level of AI ability, with a very high level of mastery of language and visual material, and more and more capabilities in a broader variety of cognitive tasks. See also "the report" for a lot more evidence, including about the disagreements on the actual current abilities. Finally, there is no scientific reason to believe that humanity is at the pinnacle of intelligence: In fact, in many specialized cognitive tasks, computers already surpass humans. Hence, even ASI is plausible (although at what level, it cannot be fathomed), and, unless one relies on arguments based on personal beliefs rather than science, the possibility of AGI and ASI cannot be ruled out.




*Performance of AI models on various benchmarks from 2000 to 2024, including computer vision (MNIST, ImageNet), speech recognition (Switchboard), natural language understanding (SQuAD 1.1, MNLU, GLUE), general language model evaluation (MMLU, Big Bench, and GPQA), and mathematical reasoning (MATH). Many models surpass human-level performance (black solid line) by 2024. Kiela, D., Thrush, T., Ethayarajh, K., & Singh, A. (2023) 'Plotting Progress in AI'.*

## For those who think AGI is possible but only in many decades

Another common argument is that it is not necessary to regulate against the risks of AGI since it has not yet been reached and it's impossible to know what it will look like. I find this argument unconvincing for two reasons: First, it is impossible to be sure that AGI will not be achieved by adding some trick on top of current methods, and the capability trend lines continue to point towards AGI. Second, and most importantly, the moment when AGI will emerge is unknown, while legislation, regulatory bodies and treaties require many years, if not decades, to be put in place. In addition, who could state with honesty and true humility that advances are certainly not around the corner? I do agree that when we compare the current most advanced general-purpose AI systems and human intelligence, we see gaps that may require new breakthroughs in AI research to be closed. In particular, I agree that current dialogue systems do not reason and plan as well as humans and display much inconsistent behavior.

But we already have systems that can reason and plan better than humans, e.g.,

with AlphaGo, albeit their knowledge is limited and was not learned but hard-coded (the rules of Go for example). So the required breakthrough would bring together the knowledge acquisition and linguistic skills of GPT-4 with the planning ability of AlphaGo. Besides, many humans also do not reason that well and can "hallucinate" answers that are not grounded in reality, or act inconsistently, both being well-studied weaknesses of LLMs, so we may not be as far as some think from the spectrum of human-level capabilities. More crucially, before ChatGPT, most AI researchers including myself did not expect its level of capabilities to arise before decades, and the three most cited experts in the field of AI are now worried of what this could mean. Given this uncertainty, I recommend we keep our beliefs open: advances could continue at the same rate, or they could stall and it could take decades to reach AGI. The only rational stance compatible with all this evidence is humility and planning with that uncertainty.

A pattern that I have sometimes observed in discussions that I find misleading is to talk as though AI capabilities will just forever remain at the current level: We do need to consider plausible future scenarios and trajectories of AI advances in order to prepare against the most dangerous ones, and take stock of the trends like in the above figure.

### For those who think that we may reach AGI but not ASI

Some believe that humans are already at the peak of possible intelligence and that AI systems will not be able to match all of our abilities. This cannot be disproved but is very unlikely, as I argued above in the first named section and as Geoff Hinton **eloquently argued** by comparing the abilities of analog computation (like in our brain) versus digital computation (like in computers). Furthermore, no need to cover all human abilities to unlock dangerous existential risk (x-risk) scenarios: It suffices to build AI systems that match the top human abilities in terms of AI research. A single trained AI with this ability would provide hundreds of thousands of instances able to work uninterruptedly (just like a single GPT-4 can serve millions of people in parallel because inference can be trivially parallelized), immediately multiplying the AI research workforce by a large multiple (possibly all within a single corporation). This would likely accelerate AI capabilities by leaps and bounds, in a direction with lots of unknown unknowns as we move possibly in a matter of months from AGI to ASI. This larger AI research workforce could construct more capable AI and further accelerate the advances, etc. Along similar lines, there is an argument that robotics currently lags quite significantly compared to more cognitive abilities of AI. Again, looking at the trend and current state, we see that improvements in robotics continue, and they could be accelerated by AGI

and ASI because an AGI with the skills of a robotics researcher would accelerate these advances. These advances should definitely be monitored closely as we can imagine that self-preserving AI systems that would not need humans anymore because they could control robots for achieving physical work would theoretically have a clear incentive to get rid of humanity altogether to rule out the possibility of humans turning them off.

### For those who think that AGI and ASI will be kind to us

I really wish that these expectations will be realized, but the clues from computer science research in AI safety point in the opposite direction and, in the absence of a clear case, risk management demands to take precautions against the plausible bad outcomes. **An AI with a self-preservation goal would resist being turned off** and in order to minimize the probability of being turned off, a plausible strategy would be for it to control us or get rid of us to make sure we would not jeopardize its future. Deals between entities that have to negotiate a mutually beneficial outcome (like between people or countries) only work when none of the sides can defeat the other with high enough certainty. With ASI, this kind of equilibrium of forces is far from certain. But why would an AI have a strong self-preservation goal? As I keep saying, this could simply be the gift made by a small minority of humans who would welcome AI overlords, maybe because they value intelligence over humanity. In addition, a number of technical arguments (around instrumental goals or reward tampering) suggest that such objectives could emerge as side-effects or innocuous human-given goals (see “the report” and the vast literature cited there, as well as the diversity of views about loss of control that illustrate the scientific uncertainty about this question). It would be a mistake to think that future AI systems will necessarily be just like us, with the same base instincts. We do not know that for sure, and the way we currently design them (as reward maximizers for example) point in a very different direction. See the next point below for more arguments. These systems may be similar to humans in some ways and very different in others that are hard to anticipate. In addition, in a conflict between two groups of humans, if one group has vastly superior technology (like Europeans invading the Americas, particularly in the 19th and 20th centuries), the outcome can be catastrophic for the weaker group. Similarly, in a conflict between ASI and humanity, our prospects could be dire.

### For those who think that corporations will only design well-behaving AIs and existing laws are sufficient

Why would engineers in the corporations designing future advanced AI systems

not only design a safe type of AI? Shouldn't corporations prefer safe AI? The problem comes when safety and profit maximization or company culture ("move fast and break things") are not aligned. There is lots of historical evidence (think about fossil fuel companies and the climate, or drug companies before the FDA, e.g., with thalidomide, etc) and research in economics showing that profit maximization can yield corporate behavior that is at odds with the public interest. Because the uncertainty about future risks is so large, it is easy for a group of developers to convince themselves that they will find a sufficient solution to the AI safety problem (see also my discussion of psychological factors in a future blog post).

Avoiding the effects of the conflict of interest between the externality of global risks and corporate interests or individual wishful thinking is why we have laws, but teams of lawyers can find legal loopholes. One can think of an ASI that is even smarter than the best legal team. It is very likely that it would find such loopholes, both in the law and in the instructions we provide to demand that the AI behavior yields no harm. The difficulty of drafting a contract that constrains the behavior of an agent (human, corporate or AI) as intended by another agent is **generally intractable**. In addition, note how we keep patching our laws in reaction to the loopholes found by corporations: It is not clear that we will be able to iterate many times with the loopholes found by an ASI. I see this problem boiling down to our inability to provide to the AI a formal and complete specification of what is unacceptable. Instead we provide an approximate safety specification *S*, presumably in natural language. When the AI is given a main goal *G* under the constraint to satisfy *S*, if achieving *G* without violating all the interpretations of *S* is easy, then everything works well. But if it is difficult to achieve both, then it requires a kind of optimization (like teams of lawyers finding a way to maximize profit while respecting the letter of the law) and this optimization is likely to find loopholes or interpretations of *S* that satisfy the letter but not the spirit of our laws and instructions. Examples of such loopholes have already been studied in the AI safety literature and include behaviors such as reward tampering (taking control of the reward mechanism and then trying to keep it creates an implicit self-preservation goal) as well as the numerous instrumental goals, where to achieve the main apparently innocuous goals, it is useful for the AI to also achieve potentially dangerous subgoals, such as self-preservation or having more control and power over its environment (e.g., via persuasion, deception and cyberhacking) and evidence of these inclinations have already been detected. What complicates matters is that the engineers do not directly design the AI's behavior, only how it learns. As a result, what it learns, at least with deep learning, is extremely complex and opaque and makes it difficult to detect and rule out unseen intentions and

deception. See “the report” for lots of references as well as pointers to AI safety research aiming to mitigate these risks, but not yet having achieved this.

### For those who think that we should accelerate AI capabilities research and not delay benefits of AGI

The core argument is that future advances in AI are thought to be likely to bring amazing benefits to humanity and that slowing down AI capabilities research would be equivalent to forfeiting extraordinary economic and social growth. That is well possible, but in any rational decision-making process, one has to put in the balance both the pros and the cons of any choice. If we achieve medical breakthroughs that double our life expectancy quickly but we take the risk of all dying or losing our freedom and democracy, then the accelerationist bet is not worth much. Instead, it may be worthwhile to slow down, find the cure for cancer a bit later, and invest wisely in research to make sure we can appropriately control those risks while reaping any global benefits. In many cases, these accelerationist arguments come from extremely rich individuals and corporate tech lobbies with a vested financial interest in maximizing profitability in the short term. From their rational point of view, AI risks are an economic externality, i.e., whose cost is unfortunately borne by everyone. This is a familiar situation that we have seen with corporations taking risks (such as the climate risk with fossil fuels, or the risk of horrible side effects of drugs like thalidomide) because it was still profitable for them to ignore these collective costs. However, from the point of view of ordinary citizens and of public policy, the prudent approach into AGI is clearly preferable when adding up the risks and benefits. There is a possible path where we invest sufficiently in AI safety, regulation and treaties in order to control the misuse and loss-of-control risks *and* reap the benefits of AI. This is the consensus out of the **2023 AI Safety Summit** (bringing 30 countries together) in the UK as well as the 2024 follow-up in Seoul and the **G7 Hiroshima principles** regarding AI, not to mention numerous other intergovernmental declarations and proposed legislation in the UN, the EU and elsewhere.

### For those concerned that talking about catastrophic risks will hurt efforts to mitigate short-term human-rights issues with AI

I have been asked to stop talking about the catastrophic risks of AI (both catastrophic misuse and catastrophic loss of control) because that discussion would suck all the air out of the room, grabbing all of the public attention, at the expense of addressing the well-established harms to human rights already happening with AI. In a democracy, we collectively have many discussions and it is

unusual for one to say, e.g., to "stop talking about climate change" by fear that it would harm discussion on child labour exploitation, or to stop talking about the need to mitigate the long-term effects of climate change because it would harm discussion on the need for short-term adaptation to that change. If the people telling me to avoid voicing my concerns had a strong argument about the impossibility of catastrophic AI risks, I would understand the undesirable possibility of introducing noise in the public discourse. But the reality is that (a) there are plausible arguments why a superhuman AI may have a self-preservation goal that would endanger us (the simplest being that humans provide it), (b) the stakes (if that danger materializes) are so high that even if it were a low-probability event, it should rationally demand our attention and (c) we do not know what is the timeline to AGI, and credible voices from inside the frontier AI labs claim that it could be just a few years, so it may not be so long-term after all, while putting up legislation and regulation or even treaties could take much more time. Our future well-being as well as our ability to control our future (or, in other words, our liberty) are human rights that also need to be defended. Also, the interests of those worried about short-term and long-term risks should converge because both groups want government intervention to protect the public, which means some form of regulation and public oversight of AI. Recent legislative proposals regarding AI tend to cover both short-term and long-term risks. In practice, those who oppose regulation are often those who have a financial or personal interest in (blindly) accelerating the race towards AGI. The tech lobby has successfully deflected or watered down attempts at legislation in many countries, and all those who ask for regulation with effective teeth should rationally unite. Sadly, this infighting among those who want to protect the public greatly decreases the chances of bringing up public scrutiny and the common good into how AI is developed and deployed.

### For those concerned with the US-China cold war

China is the second AI superpower after the US and the geopolitical conflict between China and the US (and its allies) creates a genuine concern in the Western democracies. Some believe that China could exploit advances in AI, especially as we approach AGI and ASI, that could be turned into powerful weapons. That could give China the upper hand both economically and militarily, especially if the West slows down its march towards AGI in favour of safety. First, to be fair, it is also clear that the Chinese also fear that the US could use advances in AI against them, and it motivates the Chinese government to also accelerate AI capabilities research. For those like me, who believe that democratic institutions are much better than autocratic regimes at protecting human rights (please read the **UN Universal Declaration of Human Rights**, signed by China but unfortunately not

binding), this geopolitical competition is especially concerning and presents a tragic dilemma. In particular, we already see current AI being used to influence public opinion (e.g. with deep fakes) and undermine democratic institutions by stoking distrust and confusion. Autocratic governments are already using AI and social media to solidify their internal propaganda and control dissent (including with both internet surveillance and visual surveillance through face recognition). There is therefore a risk that AI, especially AGI, could help autocrats stay in power and increase their dominance, even bring about an autocratic world government. The possibility that future AI advances could provide first-strike offensive weapons (including in the context of the cyberwar) motivates many in the West to accelerate AI capabilities and reject the option of slowing down in favour of increased safety, by fear that it would allow China to leap ahead of the US in AI. However, how do we also avoid the existential risk of losing control to an ASI, if we ignore AI safety and just focus on AI capabilities? If humanity loses because of uncontrolled ASI, it won't matter what kind of political system one prefers. We all lose. We are in the same boat when it comes to x-risk. Hopefully, that motivates leaders on both sides to look for a path where we also invest in AI safety, and we could even collaborate on research that would increase safety, especially if that research in itself does not increase capabilities. No one would want the other side to make a globally catastrophic mistake in the development of their AGI research, because a rogue ASI would not respect any border. In terms of investment, it does not have to be an either-or choice between capability and safety research, if the efforts start now. We collectively have enough resources to do both, especially if the right incentives are put in place. However, sufficient investment in safety is necessary to ensure the safety answers are figured out *before* reaching AGI, whatever its timeline, and it is not currently what is happening. I am concerned that if sufficiently safe AI methodologies are not found by that time, the more concrete risk of the adversary's supremacy may trump the existential risk of loss of control because the latter would be considered speculative (whereas the former is more familiar and anchored in centuries of armed conflicts).

### For those who think that international treaties will not work

It is true that international treaties are challenging, but there is historical evidence that they can happen or at least this history can help understand why they sometimes fail (the history of the **Baruch plan** is particularly interesting since the US was proposing to share nuclear weapons R&D with the USSR). Even if it is not sure that they would work, they seem like an important avenue to explore to avoid a globally catastrophic outcome. Two of conditions for success are (a) a common interest in the treaty (here, avoiding humanity's extinction) and (b) compliance

verifiability. The former requires governments to truly understand the risks, and more research is thus needed to better analyze it, so syntheses of the AI safety science like in "the report" will help. However, (b) is a particular problem for AI, which is mostly software, i.e., easy to modify and hide, making mistrust win against a treaty that would effectively prevent dangerous risks from being taken. However, there has been a flurry of discussions about the possibility of **hardware-enabled governance mechanisms**, by which high-end chips enabling AGI training could not be hidden and would only allow code that has been approved by a mutually chosen authority. The AI high-end chip supply chain has very few players currently, giving governments a possible handle on these chips. See also the hardware design proposed in **this memo**. One can also think of scenarios where hardware-enabled governance could fail, e.g., if ways of reducing the computational cost of training AI by many orders of magnitude are discovered. This is possible but far from certain, and none of the tools proposed to mitigate AI catastrophic risk is a silver bullet: What is needed is "defense in depth", layering many mitigation methods in ways that defend against many possible scenarios. Importantly, hardware-enabled governance is not sufficient if the code and weights of the AGI systems are not secured (since using or fine-tuning such models is cheap and does not require high-end chips or the latest ones), and this is an area which there is a lot of agreement outside of the leading AGI labs (which do not have a strong culture of security) that a rapid transition towards very strong cyber and physical security is necessary as AGI is approached. Finally, treaties are not just about the US and China: In the longer term, safety against catastrophic misuse and loss of control requires all the countries on-board. But why would the Global South countries sign such a treaty? The obvious answer I see is that such a treaty must include that AI is not used as a domination tool, including economically, and that its scientific, technological and economic benefits must be shared globally.

### For those who think the genie is out of the bottle and we should just let go and avoid regulation

The genie is possibly out of the bottle: Most of the scientific principles required to reach AGI may have already been found. Clearly, large amounts of capital is being invested with that assumption. Even if that were true, it would not necessarily mean that we collectively should let the forces of market competition and geopolitical competition be the only drivers of change. We still have individual and collective agency to move the needle towards a safer and more democratic world. The argument that regulation would fail is similarly wrong. Even if regulating AI is not going to be easy, it does not mean that efforts should not be made to design institutions that can protect human rights, democracy, and the future of humanity,

even if it means that institutional innovation is needed. And even just reducing the probability of catastrophes would be a win. There is no need to wait for the silver bullet to start moving the needle positively. For example, to deal with the challenge for states to build up the required technical capacity and have the required innovation ability, regulators could rely on private non-profit organizations that compete with each other to design more effective capability evaluations and other safety guardrails. To deal with the fast pace of change and unknown unknowns of future AI systems, rigid regulations would not be very effective, but we also have examples of principle-based legislations that provide enough freedom to the regulator to adapt to changing circumstances or new risks (think about the FAA in the US, for example). To deal with conflicts of interest (between public good and profit maximization) within corporate AI labs, the government could force these companies to have multiple stakeholders on their board that represent the necessary diversity of views and interests, including from civil society, independent scientists and the international community.

### For those who think that open source AGI code and weights are the solution

It is true that open science and open source have delivered great progress in the past and will continue to do so in general. However, one should always weigh the pros and cons of decisions like the one to publicly share the code and parameters of a trained AI system, particularly as capabilities advance and reach human-level or beyond. Open-sourcing of AI systems may plausibly currently be more beneficial than detrimental to safety because they enable AI safety research in academia while current systems are apparently not yet powerful enough to be catastrophically dangerous in bad hands or with loss of control. But in the future, who should decide where to draw the line and how to weigh the pros and cons? CEOs of companies or democratically chosen governments? The answer should be obvious if you believe in democracy. There is a difficult question (and one that is painful for me): Is freely shared knowledge always a globally good thing? If we had the DNA sequence of an extremely dangerous virus, would it be best to share it publicly or not? If the answer is obvious to you in this case, think twice about the case for AGI algorithms and parameters. A red flag came up recently: an **EPFL study** showing superior persuasion abilities of GPT-4 (compared with ordinary humans) when given Facebook pages of the person to be persuaded. What if such an AI system was further fine-tuned on millions of interactions teaching the AI how to be really efficient to make us change our mind on anything. The success of demagogy clearly shows a human Achilles' heel there. Regarding x-risk, some have made the argument that if everyone had their own AGI, the "good AIs" would win over the "bad AIs" because there are more good people. There are many flaws in

this argument. First, we are not sure at all that the goodwill of the owner of an AGI will be sufficient to guarantee the moral behavior of this AGI (see above about instrumental goals). Second, it is not at all sure that a minority of rogue AIs would be defeated by a majority of good AIs and that we will discover appropriate countermeasures in time (although we should definitely try). It depends on the defense-offense balance: think about lethal first strikes. The rogue AIs could choose their attack vector to give a strong advantage to the attacker. A prime candidate for this is the class of bioweapons. They can be developed in silence and released at once, exponentially creating death and havoc while the defender struggles to find a cure. The main reason why bioweapons are not often used in human wars is because it is difficult for the attacker to make sure their weapon will not turn against them, because we are all humans, and furthermore, even if they had a cure, once a pathogen is out, it will mutate and all guarantees could be lost. But a rogue AI intent on eliminating humanity would not be subject to this concern. Regarding the misuse of open source AI systems, it is true that even closed-source systems can be abused, e.g., with jailbreak, but it is also true that it is (a) much easier to find attacks against open source systems and (b) once released, an open source system cannot be fixed against newly discovered vulnerabilities, unlike a closed-source system. Importantly, that includes fine-tuning of the open source system that would reveal dangerous capabilities from the point of view of loss of control. An argument in favor of open source is the greater access to more people. That is true but it still takes technical expertise to fine-tune these systems and the exponentially growing computational cost of the most advanced AI systems means that it is likely that the number of organizations able to train them will be very limited, giving them immense power. I would favor a form of decentralization of power that addresses that concentration of power and does not increase the risks of misuse and loss of control, on the contrary: organizations building these systems could be forced to adopt strong multistakeholder governance, increasing transparency (of at least the capabilities, not necessarily the list of key ingredients for success) and public scrutiny to reduce the risks of abuse of the AGI power and loss-of-control mistakes because of insufficient safety methodologies. In addition, trustworthy researchers could be given controlled access to the code with technical methods preventing them from taking the code back with them, providing a greater depth of oversight and mitigation of power abuse.

### For those who think worrying about AGI is falling for Pascal's wager

Pascal's wager is that given the infinite losses (hell vs paradise) incurred if we wrongly choose to not believe in God, we should act (wager) under the belief that God (the Christian god, by the way) exists. The argument against doing something

about the catastrophic risks of AI draws the analogy to Pascal's wager because of the huge risks, even potentially infinite if you consider the extinction of humanity that way. In the limit of infinite losses under extinction, we would have to act as if those risks are real with an amount of evidence or a probability of extinction that is allowed to go to zero (because the risk can be measured, in expectation, by the product of the probability of the event times the loss if it happens). Let us now see where that argument breaks down, mostly because we are not dealing with tiny probabilities. In a December 2023 **survey**, the median AI (not safety) researcher put 5% on AI causing extinction-level harm. Probabilities of 5% are not Pascal's wager territory. There are serious arguments supported in the scientific literature (see "the report" and a lot of the discussion above) for the various kinds of catastrophic risks associated with very advanced AI, especially as we approach or surpass human level in some domains. Also, we do not need to take the losses to infinity: There are many potentially very harmful possibilities along the path to AGI and beyond (again, see "the report"). So we end up with non-zero evidence for AI catastrophes and the possibility of non-infinite but unacceptable losses, the usual setting for decision theory, and rationality thus demands that we pay attention to these risks and try to understand and mitigate them.

### For those who discard x-risk for lack of reliable quantifiable predictions

Of course no one has quantitative models of future scientific advances along with social and political change regarding AI. Hence we cannot run quantitative models such as those applied to sample future climates. The only quantitative options are individual and aggregated subjective probabilities, e.g., from **polling experts**. Can we trust the 5% median x-risk in this recent study? I would say to some extent, in ways similar to how we can trust the aggregate long-term predictions of economists. Are they sufficient to drive policy? No, not alone, but they surely send an important signal because experts internalize their understanding of the world and apply their system 1 computation to it, i.e., intuition, in ways that can be very valuable. But of course, and very importantly, we also need to consider rational but not fully quantitative arguments such as those I outlined above. For example, we can ask questions like, "what if we build a superintelligent AI, and what if it has goals that are dangerous to humanity?". There are many ways in which one can argue that superintelligence is plausible (with lots of uncertainty about the timeline) and many ways that have been discussed in which an AI acquires dangerous goals, the simplest being that a human provides them. Should this uncertainty make one conclude that public policy should not consider AI x-risk? Of course not; given the magnitude of the potentially negative impact (up to human extinction), it is

imperative to invest more in both understanding *and quantifying* the risks and developing mitigating solutions. And the uncertainty in timeline means that, yes, there is urgency in doing these things, in case AGI happens faster than expected.

Published in **Uncategorized**

Previous Post
**The International Scientific Report on the Safety of Advanced AI**

Next Post
**Bounding the probability of harm from an AI to create a guardrail**

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, and making him the computer scientist with the largest number of citations and h-index.

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

He received numerous awards, including the prestigious Killam Prize and Herzberg Gold medal in Canada, CIFAR’s AI Chair, Spain’s Princess of Asturias Award, the VinFuture Prize and he is a Fellow of both the Royal Society of London and Canada, Knight of the Legion of Honor of France, Officer of the Order of Canada, Member of the UN’s Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology. Yoshua Bengio

was named in 2024 one of TIME's magazine 100 most influtential people in the world.

Concerned about the social impact of AI, he actively contributed to the Montreal Declaration for the Responsible Development of Artificial Intelligence and currently chairs the International Scientific Report on the Safety of Advanced AI.

## Search

Search for... Search

## Recent Posts

- Implications of Artificial General Intelligence on National and International Security
- Bounding the probability of harm from an AI to create a guardrail
- Reasoning through arguments against taking AI safety seriously
- The International Scientific Report on the Safety of Advanced AI
- Towards a Cautious Scientist AI with Convergent Safety Bounds

## Categories

- AI for Social Good
- AI safety
- Climate change
- COVID-19
- Publication
- Uncategorized