# Advanced AI and the ethics of risking everything

Olle Häggström

August 6, 2025

## 1. Introduction

Imagine sitting in the back seat of a taxi with your two closest family members. As you approach the bridge across the canal, you notice that a ship is waiting to pass and that the bridge has slowly begun to lift. The driver, visibly annoyed by the delay, turns to you and your loved ones and asks “Shall we try and jump it?”. You quickly calculate that doing so would save five minutes, at the cost of a 1% probability of a crash that kills everyone in the car. Do you give the driver your permission?

In a Hollywood movie, you would probably say yes. But this is real life and you are not insane, so of course you politely decline the driver’s reckless suggestion.

Next consider Sam Altman, CEO of OpenAI, facing the decision of whether to release the newly developed GPT-5. (There’s a good chance that when this reaches the reader, release of GPT-5 has already happened, but at the time of writing, in August 2025, it is still a hypothetical future event.) The situation has some similarities to the case of the taxi and the bridge opening. As in that case, the action under consideration is meant to have a benign consequence: here, to make a highly useful product available to many millions of customers across the world. And again as in the bridge jumping case, the hope for this benign outcome needs to be weighed against the risk of a very bad one: here, the release of an AI that is both capable of and motivationally inclined towards wresting control of the world from humans and possibly even killing everyone. Also similar is the sharp asymmetry between the hoped-for benign effect and the risk side, as in both cases the badness that results if things go wrong is many orders of magnitude greater than the goodness of the intended outcome.

There are also some glaring disanalogies between the two situations. One is that while the bridge opening thought experiment assumes that the probability of catastrophe is known, there are no reliable such probability estimates in the case of releasing new AI models. Also, the stakes in the AI case are vastly greater than in the bridge opening case: the five minutes saved in the taxi ride is insignificant compared to the total amount of good that GPT-5 might do, while the four lives at stake in the taxi is very little compared to the eight billion human lives (plus the rest of the biosphere) that might perish if the AI goes rogue. Yet another difference is that while the taxi driver has the courtesy to ask the people whose lives he may be about to risk for their permission to do so, Sam Altman has asked very few or perhaps none of his eight billion fellow human beings for permission to risk their lives. Although as we shall see there are plenty of further complications, it seems that all three of these differences point in the same direction, namely that Altman’s obligation to show restraint and to avoid the risky choice is even greater than that of the taxi driver.

This paper is about the ethics of the choice that Altman is facing regarding the possible release of GPT-5, and more generally of the leading AI companies’ choice to push full speed ahead on developing and releasing ever more advanced AI despite the catastrophic risk involved. The rest of the paper is organized as follows. For the sake of those readers who are not yet onboard with the kind of catastrophic risk discussed here being a real thing, I will recall in Section 2 some of the basic arguments for why it is. Then, in Section 3, I will discuss attempts at quantifying what risk levels are tolerable, followed in Section 4 by a short discussion of the role of consent. Section 5 deals with the safety testing procedures that the leading AI companies are subjecting their models to pre-deployment, and why fundamentally they do not work. Having said by that point more than enough to conclude that the risk that these companies are exposing all of us to is unacceptable, I move on in Section 6 to the incentives and the psychology that causes them to nevertheless do so, and finally in Section 7 I offer some concluding remarks.

Before embarking on that agenda, let me note that the choice to risk literally everything on the planet is not entirely new to the AI era, but does have one striking predecessor. As the 1940s Manhattan project approached the stage where it was ready for the first actual nuclear detonation test, there was some lingering uncertainty as to whether such a detonation might cause an uncontrolled chain reaction igniting the atmosphere and ending life on Earth. Probably not, was the consensus view among the participating physicists, but due to the high stakes three of them were assigned to look closer into the problem and hopefully achieve more clarity. The resulting report ends with a paragraph summarizing the expected finding – that atmospheric ignition seemed very unlikely – but the final sentence is less than reassuring: “However, the complexity of the argument and the absence of satisfactory experimental foundations makes further work on the subject highly desirable” (Konopinski et al, 1946). Despite this remaining inconclusiveness, the project went on with the Trinity detonation in July 1945. The atmosphere did not ignite, and we now know that it couldn’t happen, but the point is that those in charge did not know that.

Fast forward to April 2023 and OpenAI’s release of GPT-4. The accompanying technical report contains a safety section with, among other things, a discussion of whether the model might be able to ”autonomously replicate and gather resources” – two of the most central competences in various catastrophic AI scenarios. It concludes that the model “is probably not yet capable of” doing so, but cautions, in a formulation eerily similar to that of the Konopinski report, that “further research is needed to fully characterize these risks” (OpenAI, 2023a). And yet, as with the Trinity test, GPT-4 was released.

## 2. Basics of AI existential risk

The rest of this paper is premised on the idea that existential risk from AI is a valid concern. Here there is only room to deal with this idea very sketchily, so readers unfamiliar with the literature on this would do well to read one of the classic masterpieces – Bostrom (2014) or Russell (2019) – and to complement this with one or two more up-to-date surveys such as Hendrycks et al (2023), Leahy et al (2024) or Häggström (2025). There is good reason to believe that the upcoming book by Yudkowsky and Soares (2025) is an excellent alternative recommendation, but I haven’t yet had access to it. For a concrete and extraordinarily well-researched scenario where things go astray already in the present decade, see the *AI 2027* report by Kokotajlo et al (2025).

Borrowing from Häggström (2024), the case for the reality and urgency of AI existential risk can be structured as follows.

> (1) **AI capabilities.**
>
> > (1a) It is possible to build a machine which, in terms of cognitive capabilities relevant to the ability to take over the world, is vastly superior to a human.
> >
> > (1b) This is achievable in the not-too-distant future.
>
> (2) **AI motivations.** Unless we make sure that such an AI's goals are highly aligned with human values, or whatever values it is we want it to pursue to benefit humans, this AI may develop very different goals, and go on to prioritize those over human welfare.

If we accept these claims, it pretty much follows that if we continue to build ever more capable AIs without solving the AI alignment problem indicated in (2), we risk creating superintelligent AIs with goals alien to ours, leading to potential conflict and possibly even the end of humanity.

A common way to try to evade this conclusion is to insist that as long as advanced robotics lags behind, AIs will essentially be restricted to producing text in text windows, so nothing bad can happen in the physical world. This overlooks two things: first, the strong current trend towards AI agents with Internet tool access (see, e.g., Docker and Mowshowitz, 2025), and second, that an AI without access to robots may instead use social manipulation skills to employ individual humans for doing its biddings out there. More on the latter in Section 5.

So what about Claims (1a), (1b) and (2)? Claim (1a) pretty much follows from two very natural assumptions: that humans as products of imperfect biological evolution are nowhere near the ceiling of attainable intelligence levels, and that cognition is at its core an information processing phenomenon that does not require a biological substrate but can be implemented *in silico*. When it comes to the current generative AI paradigm, attempts are sometimes made to point to some property of these AIs that show a complete lack of "real" (whatever that means) intelligence, but such arguments are typically easily adapted so as to show that humans lack intelligence as well, thereby making a compelling case that the original argument was flawed; see Häggström (2023) for a series of examples.

Claim (1b) about the timing of the emergence of superintelligence is much more open to reasonable debate. Triggered by the very rapid AI development over the last 5-10 years, experts' timelines have tended to shrink drastically, and today many of these are measured in single-digit years rather than in decades – Kokotajlo et al (2025) being a prime example.

Central to parts of the discussion is the idea of an intelligence explosion driven by a spiral of recursive AI self-improvement. The notion goes back quite far (Good, 1965; Solomonoff, 1985), but remained for a long time very idealized, and it is only recently that it has become better grounded in empirical observation (Aschenbrenner, 2024; Eth and Davidson, 2025; Kokotajlo et al, 2025). But even if we accept that an intelligence explosion may take off very quickly – which is still uncertain – there is the issue of how long it will be until the most efficient AI researchers are not flesh-and-blood humans but AIs, so that the recursion can take off. Instrumental in making analyses of this question go from mostly handwaving to more rigorously empirical is the work of Kwa et al (2025) and METR (2025b), demonstrating how language model capabilities to correctly complete difficult tasks (as measured by how long it takes human experts to do them) have improved exponentially with doubling times so short that extending the trends just a few years into the futures leads to astounding consequences.

Finally, there is Claim (2) on what a superintelligent AI would be inclined to do, absent successful AI alignment. Simplifying somewhat, our present-day understanding of this issue can be said rest mainly on two bodies of work. The first consists of the classical works of Omohundro (2008), Yudkowsky (2008), Bostrom (2014) and others on the so-called Orthogonality and Instrumental Convergence Theses, where the former states that pretty much any ultimate AI goal (including outrageous sounding ones such as paperclip production maximization) is compatible with arbitrarily high intelligence levels, and the latter identifies a number of instrumental goals (including self-preservation and resource acquisition) that a sufficiently capable AI is likely to form almost regardless of its ultimate goal. The second, and rapidly growing, body of work consists of empirical findings of how present-day AI systems exhibit egregiously misaligned behavior, despite the developers' efforts to eradicate such behavior; see, e.g., Greenblatt et al (2024) and Anthropic (2025) for some laboratory findings, as well as Piper (2025) and Pressman (2025) for recent examples in the wild. For concrete examples of what such misalignment might entail for increasingly capable systems, I refer once again to Kokotajlo et al (2025).

### 3. The Faust parameter

We live in an inherently dangerous world, and there is no way we can bring risk down to zero. Suppose we change the thought experiment I began with in Section 1 by having the taxi driver not suggest jumping a bridge, but instead some less dangerous traffic maneuver whose probability of resulting in a fatal accident is not $10^{-2}$ but $10^{-30}$. Then I would not object, because that is such an insignificant probability – far smaller than the risk of death I expose myself to every time I cross a street or do pretty much anything else in traffic.

In fact, whenever I cross a street, even the risk of initiating a chain of events that causes global catastrophe is nonzero, and I daresay greater than $10^{-30}$; such events may not be super common, but they do happen, as exemplified by the wrong turns and the chance meetings in Sarajevo 1914 that led to the start of World War I. And if GPT-5 similarly has probability $10^{-30}$ of ending human civilization, then that is hardly worth raising an eyebrow over.

So presumably we can draw a line somewhere above $10^{-30}$ – a line indicating the maximum probability of GPT-5 causing existential catastrophe that we find acceptable. (Note that there are many other potentially valid reasons why we might deem the release of GPT-5 unacceptable, such as various more mundane risks; my zooming in here on existential risk is not meant to downplay these other reasons, but just to keep the discussion focused.) But where should this line be drawn? Would $10^{-6}$ (one in a million) be an acceptable probability? A straightforward expected value calculation might suggest that the answer is no: if GPT-5 has probability $10^{-6}$ of killing everyone, i.e., $8 \cdot 10^9$ people, then on average it kills $8 \cdot 10^9 \cdot 10^{-6} = 8000$ people, which is an awful lot of corpses to build a single technical product on. But here we are on shaky grounds – in the vicinity of Blaise Pascal's famous Wager – for doing the expected value calculation, as we are multiplying a very small probability with a very large consequence, a practice whose validity has been debated endlessly among decision theorists and others; see, e.g., Weitzman (2009), Bostrom (2009), Häggström (2016), Kosonen (2022) and Beckstead and Thomas (2024). So while an expected value calculation landing in 8000 dead people needs to be given very serious weight in decisions involving such risks, I am, in the light of the continuing struggles to make sense of almost-Pascalian decision theory, not prepared to dogmatically declare taking such risks morally prohibited solely based on the expected value calculation, or to quantify exactly how large the upside must be to make the risk worth taking.

One may attempt to boost the expected value argument against risking everything by noting that killing everyone also obliterates the possibility of all future generations; this is the approach of Parfit (1984) and Bostrom (2013), and has become a core tenet among subsequent thinkers who label themselves *longtermists* (MacAskill, 2022; Greaves et al, 2025). Under conservative assumptions, Bostrom estimates the potential number of future humans on this planet as $10^{16}$, and replacing the current world population by that number in the above calculation gives an expected death toll of not 8000 but $10^{16}\cdot10^{-6}$ = 10 billion, which is of course way worse. (He also offers alternative estimates based on assumptions about mastering intergalactic space colonization and/or mind uploading – estimates so large that even with the aforementioned microscopic catastrophe probability of $10^{-30}$ some of them lead to an expected death toll in the trillions or more. Taking such calculations literally would be completely impractical when crossing streets etc, and here we're obviously way further off into suspiciously Pascalian territory.) While I do think longtermism has a lot going for it, still the inclusion of hypothetical future lives in expected value calculations as well as the ideology itself have come under some fire (see, e.g., Kuhlemann, 2019, and Torres, 2021), and in recent years I have come to consider it an impractical detour in arguing for action against AI-related existential risk. Why argue about the lives of people a million years down the line when an AI threatens to have you and me and all our loved ones along with the rest of humanity killed by 2030?

So what probability of having *Homo sapiens* wiped out by AI is acceptable? Taking this to be a matter of individual judgement, Aaronson (2023) proposes to call this number one's *Faust parameter*, and boldly declares his own to be 2%: if continued AI development leads to an extinction probability $p$, then if $p<2\%$ this is not, in his view, a good enough reason to halt AI development. To some, this number may come across as shockingly large, but Aaronson argues that even if we do not build advanced AI, there is so much else that threatens to obliterate us, and that the hope of AI saving us from those other risks is worth those 2%. I will not declare my own Faust parameter, other than to say that it's way above the aforementioned $10^{-30}$, but also way below current risk levels which I judge to be well into the realm of double-digit percentages. This last judgement can of course be contested, but the onus should be on the AI developers to convincingly demonstrate, prior to deployment, that the risk is small. Under the present risk evaluation paradigm, they have so far (and as we shall see in Section 5) utterly failed to do so.

What I will say, however, is a caution against the simple-minded use of the Faust parameter as a guide to acceptable risk levels in connection with the release of a new AI model. If we assume for the sake of argument a Faust parameter of 2%, then if Sam Altman takes this to be an acceptable risk level for GPT-5 to destroy the world, this would presumably also apply separately to GPT-6, GPT-7, and so on, along with releases of new versions of Claude, Gemini, DeepSeek and whatnot. Probabilities will then accumulate, and after a few dozen releases our species will more likely be extinct than alive.

## 4. Consent

At first glance it seems plausible that it would be immoral of Sam Altman to risk my life without my informed consent. Yet, it is not straightforward whether and how the concept of informed consent, commonplace in medicine, should be extended to large engineering projects with many third parties affected. Wong (2016) discusses this in the case of geoengineering, which among examples in the literature is perhaps the closest analogy to the deployment of superhumanly capable AI, due to the potentially enormous effects on virtually everyone on the planet. Hansson (2006) and Varelius (2008) offer more general discussions. One problem with such extensions of the informed consent concept is that giving every person the right to veto would in effect constitute a prohibition on projects affecting many people, while anything less would in Hansson's view be inappropriate to call "informed consent" because it "leaves individuals without the right to opt out". In the former case, the technological progress that over the course of history has created so much wealth would grind to a halt, whereas in the latter case some citizens would be affected without having given consent.

The bioethics notion of informed consent, therefore, does not seem well-suited to issues of societal-scale consequences of engineering projects. This is not to say, however, that it is fine for the leading AI companies to proceed with their ambitious plan without democratically legitimate collective consent from the population at large. Here we may note that even if these plans succeed and do not result in existential catastrophe, they are likely to wreak havoc in every sector of the labor market and radically transform all aspects of society (as indicated, e.g., in sketches by Amodei, 2024, and Altman, 2025). On the face of it, proceeding with this without clear democratic consent seems like an injustice against all of us. To rectify this situation, we would need democratic elections where AI risk is at the front and center of the campaigns, along with an electorate that is well-informed on this topic. At present, we have none of this.

## 5. The bankruptcy of AI evals

Under the heading of *AI evals*, leading AI developers evaluate potentially dangerous frontier model capabilities prior to deployment. Following Häggström (2025), I will here briefly describe OpenAI's first version of their so-called Preparedness Framework for this (OpenAI, 2023b), and just note that their main competitors have similar frameworks, including Anthropic (Anthropic 2023) and Google (Dragan et al, 2024).

The Preparedness Framework involves testing model capabilities in four potentially dangerous areas. The first is cybersecurity: a rogue AI that is able to walk through firewalls and roam freely across cyberspace would be terribly dangerous. The second is CBRN (Chemical, Biological, Radiological and Nuclear) which deals with the AI's ability to do meaningful work on development and deployment of weapons of mass destruction. The third is persuasion, which may be crucial to an AI's ability to take over in case it chooses to use humans rather than robots to carry out its intentions in the physical world. The fourth category is model autonomy, which includes planning, self-exfiltration, and abilities in AI R&D that might help it kick off a spiral of recursive self-improvement. In each category, approximate thresholds are defined for risk levels Low, Medium, High and Critical, and then the model's overall risk level is taken to be the maximum over the four categories. Finally, the framework involves self-imposed rules regarding what actions (such as public deployment of the model) OpenAI can and cannot take depending on this risk level.

So far so good: the AI eval approach seems to have worked, at least in the sense that none of the models deployed so far have caused society-scale disaster. That, however, is mostly because we are not yet in a situation where the models' capabilities are sufficiently advanced to fully stress-test the eval procedures. This may soon change, and present-day evals procedures will then cease to work, as is increasingly often pointed out both in the scientific literature (Cohen et al., 2024) and in popular press (Henshall, 2024).

There are several major problems with current AI evals. The most obvious one is that no finite amount of testing can meaningfully exhaust the space of situations and promptings that the models may encounter when deployed in the wild. We do not know what we are missing, but we do know at least since our first summer with GPT-4 in 2023 that frontier models tend to keep exhibiting previously undiscovered capabilities for many months after their deployment.

A second problem is that evals do not work if the models have the cleverness and the situational awareness to sandbag or otherwise deceive us during the testing phase. For the testing to make sense, the test results must be trustworthy, and therefore the models being tested must not have the ability and inclination to deceive us. Obviously, this makes the entire procedure largely circular, and therefore, strictly speaking, useless. Recent results on so-called alignment faking and related phenomena strongly suggest that we are close to the point where frontier models do have this ability; see Meinke et al (2024) and Greenblatt et al (2024). Worth noting in this context is that in OpenAI's recent update of the Preparedness Framework, sandbagging (i.e., the model pretending during testing to be less capable than it is) is categorized among capabilities that they are not testing for within the present framework, but looking into in the hope of developing future test methods for it (OpenAI, 2025); to me, this sounds very close to an explicit admission that their current approach to evals doesn't work.

A third problem, discussed by METR (2025a) and others, is that while the evals are said to be carried out pre-deployment, this is only partly true, because in order for testing to be possible, the models need to be deployed internally, to the testing team. We should not pretend that that is safe. For instance, if a model is dangerously good at persuasion and social manipulation, it would be reckless to assume that the personnel who carry out the testing are immune to such manipulation. It therefore needs to be verified, prior to the evals, that the model lacks such social manipulation capabilities, which is impossible in the current paradigm where such verification is meant to happen during the evals.

These are serious problems with the current evals approach. When it's time for the real deal – models with truly powerful capabilities – we will need better methods, but no one knows in advance when the real deal is, so the sane and conservative approach is to assume it is now. Yet, no methods that overcome these problems are currently on the table.

## 6. The psychology of risking everything

Summing up where the arguments in the preceding sections are pointing us, it is all very much towards the conclusion that the building of ever-more-powerful AIs that the leading AI companies are engaging in, without being able to provide meaningful guarantees that these AIs won't one day kill us all, is morally impermissible. Why, then, are they still doing it?

Speculating about others' psychology is a fraught and potentially uncharitable exercise, so I will be brief. Still, the leaders of the foremost companies in the AI race are among the most powerful people in the world, and it is therefore important that the rest of us understand what drives them, even when they are not fully transparent. In the case of Sam Altman, he does a fair amount of writing (such as Altman, 2025) about how he envisions the future, just like his counterpart at Anthropic, Dario Amodei (Amodei, 2024). But he has also become famous for being "not consistently candid" (see, e.g., Field, 2024), and in a recent biography he comes across as distant and opaque throughout (Hagey, 2025).

A natural suggestion when someone acts recklessly is that they might not be aware of the risk. In the present case, this hypothesis is untenable, given how much they have spoken about existential AI risk in recent years. In interviews in 2023, Altman spoke repeatedly of "lights out for all of us" as a worst-case scenario if we fail to manage the transition to a world with advanced AI well; see, e.g., Jackson (2024). Amodei has even quantified that risk as being in the range 10 to 25 percent (Bartlett and Amodei, 2023), and both Altman and Amodei appeared as co-signatories along with Demis Hassabis (head of Google DeepMind) and many other industry leaders on a much-discussed open letter in May 2023 on extinction risk from AI (Hinton et al, 2023).

So these leaders are obviously aware of the risk, and we must look for other reasons for their behavior. Some candidates here are (a) a kind of "if we don't do it, then someone else will" logic, (b) the belief within one company that it is important for humanity as a whole that they build superintelligence before other companies, (c) a kind of macho "we can handle it" company culture with respect to solving AI alignment in time, and (d) various short-term company and market incentives.

There's an Altman quote from 2019 which serves well to illustrate (a). Echoing a statement by Manhattan project leader Robert Oppenheimer decades earlier about the "profound and necessary truth that deep things in science are not found because they are useful; they are found because it was possible to find them" (Rhodes, 1987), Altman stated that "technology happens because it is possible" (Gardner, 2023).

Explanation (b) is well-supported by the history of how the companies came about, as summarized by Leahy et al (2024). DeepMind was founded explicitly on the idea of creating superintelligence for the benefit of all of humanity, while later OpenAI came about from a combination of a similar idealistic idea with distrust in DeepMind, and yet later first Anthropic and then xAI were similarly launched due to analogous distrust in OpenAI.

And of course, support for explanations (c) and (d) is easy to find in various quotes from executives and developers at these companies. I believe all four mechanisms (a)-(d) contribute to the race we are witnessing, but wish to stress here that an *explanation* for a behavior is not automatically an *excuse* for it. For instance, I think the combination of (a) and (d) amount to anti-social behavior similar to always defecting in prisoners' dilemma and tragedy-of-the-commons games, thereby inviting the kind of socially undesirable outcomes colloquially named Moloch (Alexander, 2014).

## 7. Concluding remarks

Supposing the reader agrees with me that a good case has been made that rushing ahead with AI development in the way that the leading companies are currently doing is morally impermissible, what should the rest of us do about it? We can tell these companies to stop, as I think we should, but since they are driven by incentives unaligned with the case laid out here (whether or not the explanations I suggested in Section 6 are bulls-eye), we cannot count on such a call being sufficient. It probably needs to be supplemented by fierce regulation, first in the United States (because that is where the leading AI companies are situated), and shortly after that in binding international agreements. Achieving that is of course a highly nontrivial matter, as is responding to all the various cynical or pessimistic reactions to such hopes (including the all-too-common “but China” argument, which tends to overlook that Chinese leaders are unlikely to be more eager than their American counterparts to destroy the world). This, however, falls outside the scope of the present essay.

I’ll just end by noting that polls suggest that, at least in the West, there is a silent majority in support of pulling the brakes on the dangerous race towards superintelligence; see, e.g., Samuel (2023) and Perrigo (2025). Mobilizing a sufficiently large part of this majority to create a political momentum that puts pressure both on the AI companies and on legislators and political leaders seems likely to be an important part of the solution to the terrible situation we are facing.

## References

Aaronson, S. (2023) Should GPT exist?, *Shtetl-Optimized*, February 22.

Alexander, S. (2014) Meditations on Moloch, *Slate Star Codex*, July 30.

Altman, S. (2025) *The Gentle Singularity*, https://blog.samaltman.com/the-gentle-singularity

Amodei, D. (2024) *Machines of Loving Grace: How AI Could Transform the World for the Better*, https://darioamodei.com/machines-of-loving-grace

Anthropic (2023) Anthropic's responsible scaling policy, September 19, https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Anthropic (2025) Agentic misalignment: How LLMs could be insider threats, June 21, https://www.anthropic.com/research/agentic-misalignment

Aschenbrenner, L. (2024) *Situational Awareness: The Decade Ahead*, https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

Bartlett, L. and Amodei, D. (2023) Anthropic CEO on leaving OpenAI and predictions for future of AI, *The Logan Bartlett Show*, https://www.youtube.com/watch?v=gAaCqj6j5sQ

Beckstead, N. and Thomas, T. (2024) A paradox for tiny probabilities and enormous values, *Noûs* **58**, 431-455.

Bostrom, N. (2009) Pascal’s mugging, *Analysis* **69**, 443-445.

Bostrom, N. (2013) Existential risk prevention as global priority, *Global Policy* **4**, 15-31.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Cohen, M., Kolt, N., Bengio, Y., Hadfield, G. and Russell, S. (2024) Regulating advanced artificial agents, *Science* **384**, 36-38.

Docker, G. and Mowshowitz, Z. (2025) Understanding AI agents: Time horizons, sycophancy, and future risks, *Future of Life Institute Podcast*, May 9, https://futureoflife.org/podcast/understanding-ai-agents-time-horizons-sycophancy-and-future-risks-with-zvi-mowshowitz/

Dragan, A., King, H. and Dafoe, A. (2024) Introducing the Frontier Safety Framework, Google DeepMind, May 17, https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/

Field, H. (2024) Former OpenAI board member explains why CEO Sam Altman got fired before he was rehired, *CNBC*, May 29.

Gardner, D. (2023) Technology is not “inevitable”, *PastPresentFuture* Substack, April 9.

Good, I.J. (1965) Speculations concerning the first ultraintelligent machine, *Advances in Computers* **6**, 31-88.

Greaves, H., Barrett, J. and Thorstad, D. (2025) *Essays on Longtermism: Present Action for the Distant Future*, Oxford University Press, Oxford.

Greenblatt, R. and 19 others (2024) Alignment faking in large language models, https://arxiv.org/abs/2412.14093

Hagey, K. (2025) *The Optimist: Sam Altman, OpenAI, and the Race to Invent the Future*, W.W. Norton, New York.

Hansson, S.-O. (2006) Informed consent out of context, *Journal of Business Ethics* **63**, 149-154.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2023) Are large language models intelligent? Are humans? *Computer Science and Mathematics Forum* **8**(1), 68.

Häggström, O. (2024) On the troubled relation between AI ethics and AI safety, to appear in *Contemporary Debates in the Ethics of Artificial Intelligence* (eds S. Nyholm, A. Kasirzadeh and J. Serilli), Wiley, New York, https://www.math.chalmers.se/~olleh/AIethicsVSAIsafety.pdf

Häggström, O. (2025) Our AI future and the need to stop the bear, https://www.math.chalmers.se/~olleh/AIandHumanCivilization.pdf

Hendrycks, D., Mazeika, M. and Woodside, T. (2023) An overview of catastrophic AI risks, https://arxiv.org/abs/2306.12001

Henshall, W. (2024) Nobody knows how to safety-test AI, *Time Magazine*, March 21.

Hinton, G. and 500+ co-signatories (2023), Statement on AI Risk, Center for AI Safety, May 29, https://www.safe.ai/work/statement-on-ai-risk

Jackson, S. (2024) The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us', *Business Insider*, May 29.

Kokotajlo, D, Alexander, S., Larsen, T., Lifland, E. and Dean, R. (2025) *AI 2027*, https://ai-2027.com/

Konopinski, E., Marvin, C. are Teller, E. (1946) Ignition of the atmosphere with nuclear bombs, technical report LA-602, Los Alamos National Laboratory, http://library.sciencemadness.org/lanl1_a/lib-www/la-pubs/00329010.html

Kosonen, P. (2022) *Tiny Probabilities of Vast Value*, Ph.D. thesis, University of Oxford, https://ora.ox.ac.uk/objects/uuid:822703dc-56ba-4717-98b4-663d251e8acb

Kuhlemann, K. (2019) Complexity, creeping normalcy and conceit: sexy and unsexy catastrophic risks, *Foresight* **21**, 35-52.

Kwa, T. and 24 others (2025) Measuring AI ability to complete long tasks, https://arxiv.org/abs/2503.14499

Leahy, C., Alfour, G., Scammell, C., Miotti, A. and Shimi, A. (2024) *The Compendium*, https://www.thecompendium.ai/

MacAskill, W. (2022) *What We Owe the Future*, Oneworld Publications, London.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R. and Hobbhahn, M. (2024) Frontier models are capable of in-context scheming, https://arxiv.org/abs/2412.04984

METR (2025a) AI models can be dangerous before public deployment, January 17, https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/

METR (2025b) How does time horizon vary across domains?, July 14, https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (eds P. Wang, B. Goertzel and S. Franklin), IOS, Amsterdam, 483-492.

OpenAI (2023a) GPT-4 technical report, https://arxiv.org/abs/2303.08774

OpenAI (2023b) Preparedness Framework (Beta), December 18, https://cdn.openai.com/openai-preparedness-framework-beta.pd

OpenAI (2025) Preparedness Framework: Version 2, April 15, https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Parfit, D. (1984) *Reasons and Persons*, Oxford University Press, Oxford.

Perrigo, B. (2025) The British public wants stricter AI rules than its government does, *Time*, February 6.

Piper, K. (2025) Grok's MechaHitler disaster is a preview of AI disasters to come, *Vox*, July 11.

Pressman, J.D. (2025) On "ChatGPT psychosis" and LLM sycophancy, https://minihf.com/posts/2025-07-22-on-chatgpt-psychosis-and-llm-sycophancy/

Rhodes, R. (1987) *The Making of the Atomic Bomb*, Simon & Schuster, New York.

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, New York.

Samuel, S. (2023) AI that's smarter than humans? Americans say a firm "no thank you", *Vox*, September 19.

Solomonoff, R. (1985) The time scale of artificial intelligence: Reflections on social effects, Human Systems Management **5**, 149-153.

Torres, E. (2021) Against longtermism, *Aeon*, October 19.

Varelius, J. (2008) On the prospects of collective informed consent, *Journal of Applied Philosophy* **25**, 35-44.

Weitzman, M. (2009) On modeling and interpreting the economics of catastrophic climate change, *Review of Economics and Statistics* **91**, 1-19.

Wong, P.-H. (2016) Consenting to geoengineering, *Philosophy & Technology* **29**, 173-188.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in *Global Catastrophic Risks* (eds N. Bostrom and M. Ćirković), Oxford University Press, Oxford, p 308–345.

Yudkowsky, E. and Soares, N. (2025) *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*, Little, Brown and Company, New York.