Our AI future and the need to stop the bear

Olle Häggström

February 21, 2025

1. Locating AI on the Technological Richter Scale

AI technology is currently advancing at breakneck speed. How radically can we expect it to alter our lives, society, and our entire civilization?

This is one of the questions addressed in Nate Silver's wide-ranging 2024 book *On the Edge: The Art of Risking Everything*. To give the question some structure, he invents the so-called Technological Richter Scale (TRS), meant to indicate the magnitude of the impact of a technological invention. He borrows the logarithmic scaling from the original seismological counterpart, where an increase of the magnitude of an earthquake by 1 corresponds to increasing the energy it releases by a factor of roughly 31.6 (so that an increase in magnitude by 2 multiplies the energy by $31.6^2=1000$).

The TRS scale is less precisely defined, but Silver explains that an invention with TRS magnitude 1 corresponds to "a half-formulated thought in the shower", while TRS 2 is something that the inventor goes on to implement but does not make public, such as "a slightly better method to brine a chicken that only you and your family know about". From there, the scale moves on towards greater impact in terms, e.g., of patents and commercialization, until at TRS 6 we arrive at the level of a technology that can plausibly be short-listed for a "technology of the year" award (Silver suggests the Post-it note as a low 6 and the VCR recorder as a high 6). Similarly, TRS 7 corresponds to a candidate for being the invention of the decade (with credit cards being a low 7 and social media a high 7), TRS 8 is a candidate for invention of the century (such as electricity or the Internet), and TRS 9 has the corresponding status with respect to a millennium (here Silver stresses that only a few such technologies have been invented, such as fire, the wheel and the printing press). Finally, TRS 10 is reserved for impacts so vast that they can only be compared to the onset of the Holocene, which is the geological epoch that began some 12,000 years ago and is characterized by *Homo sapiens* as the dominant driver of changes to the planet. If AI takes over from humans as the Earth's dominant force, then that's a 10.

But is that to be expected? Or if otherwise, where will AI fall on the TRS scale? We do not know, because most likely we have neither yet seen the technology come to maturity, nor its full impact on society. Silver judges, however, that it has already passed TRS level 6, in which case it must be a 7 or higher. True to the Bayesian approach that he advocates in the book, Silver does not make a firm prediction on where AI will land, but instead offers his subjectively estimated probability distribution, in which he distributes some 90% of his probability roughly equally among TRS levels 7, 8 and 9, and places the remaining 10% at TRS 10.

Fair enough, but personally I would shift some probability mass upwards on the TRS scale compared to Silver's. Mowshowitz (2024a) argues that even if AI development should happen to quickly fizzle out at or near 2024 levels, there is still so much unrealized impact baked into this technology that a 7 is beginning to look unlikely. That seems right, and furthermore we seem to be on the verge of a regime where access to powerful AI becomes the key enabler for the design and creation of even more powerful AI. Due to this feedback mechanism, it seems to me that, short of a nuclear Holocaust or other global catastrophe that puts an end to further

technological development, or a full collective decision by humanity to purposely halt AI progress, some unforeseen obstacle would be needed for AI to stop at level 8 or 9 rather than moving on to the next level. That could still happen, of course, but I am nevertheless inclined to put rather more than Silver's 10% probability at TRS 10 – perhaps even north of 50%.

In this context it is also worth taking a step back and asking what property it is that during the Holocene has enabled *Homo sapiens* to take such remarkable control of the planet. This seems to have little or nothing to do with our muscular strength or physical endurance. Rather, it is overwhelmingly a consequence of our intelligence. If we accept this, along with the observation that intelligence is a uniquely powerful resource that has allowed us to go all the way from the savannah to the Moon, then this strongly suggests that the present stage of technological development, when we are automating this resource and delegating it to machines, may turn out to be the most crucial phase ever in human history. Whether or not this transition goes all the way to a point where machines take over our roles as the most intelligent and powerful creatures on the planet (thereby making AI a TRS 10 technology), I expect the impact of AI on human civilization to be highly transformative. Defending this view in more detail will be a major theme later on in this essay.

Expecting the impact of AI technology to be large leaves open the question of whether this impact – all things considered – will be mostly good or mostly bad for humanity. Silver addresses this question as well, and is absolutely right to point out that the higher the TRS level reached by AI, the more polarized should our estimate of the value of the outcome be. In case of AI hitting TRS 10, it would be far-fetched and strange to expect that pros and cons of the technology would mostly cancel out, with the net value to us landing somewhere in the middle of the good-bad scale. Much more likely is for such AI to *either* become the crucial technology for liberating us from labor, for eliminating scarcity, disease and death, for expanding into outer space, and more generally for realizing our dreams, *or* to bring upon us the greatest catastrophe ever, possibly including human extinction. Opportunities are enormous, but so are the risks.

Focus in what follows will be mostly on the risk side of this utility spectrum. This is not to say that discussions about what we can achieve at the positive end of the spectrum if we manage to steer AI development in desirable directions are uninteresting – on the contrary, and I recommend Bostrom (2024) and Amodei (2024) for engaging treatments of this topic from the vantage points of two very different temperaments. Rather, my emphasis on the risks stems from a realization of how urgent the need to mitigate them has become. In a recent podcast, AI researcher Connor Leahy went further in his colorful motivation for this prioritization by stating that "it is not useful to philosophize about the communist utopia if a bear is currently ramming down your door" (Faggella and Leahy, 2025, 1:06:35 into the video).

The essay will therefore turn into a kind of level-headed defense of the so-called "AI doomer" position – a term that is often used about those of us who take the risk that AI causes human extinction seriously. While with that definition I readily qualify as an AI doomer, I do nevertheless have reservations about the terminology, which is designed to evoke the image of a doomsayer proclaiming "the end is near", while on the contrary most AI doomers (myself included) take pains to emphasize that we (humanity) can *avoid* the end if we get our act together.

The rest of this essay is organized as follows. In the next section, I will offer a very brief history of AI up to early 2025, and in Section 3 and 4 I will discuss where this trajectory may be leading in the next few years. Together with some supporting arguments on AI goals and motivations in Section 5, this leads to the conclusion that urgent action is needed to stop Leahy's bear and to

protect the continued survival of our species. The essay ends with a discussion in Sections 6 and 7 of what this action might consist in.

Before all that, however, let me offer a few caveats about what has been said so far:

1. My focus here on the biggest and most transformative (i.e., TRS level 10) AI consequences is in no way meant to say that more down-to-Earth issues around, e.g., AI bias, deepfakes, privacy, intellectual property and carbon footprints are unimportant. On the contrary, those issues are very important, and it is great that there are people engaged in them. See Häggström (2024b) for my take on the pernicious idea of a conflict between work on transformative AI and on more mundane AI issues.
2. While forming a subjective probability distribution on the TRS level eventually achieved by AI – and engaging in prediction more generally – is a useful exercise that I strongly encourage the reader to engage in, I want to stress that we are still in a position to influence the future. If we believe too strongly in our predictions we risk falling into the a passive mode of just waiting for them to be realized. This is true even if the predictions are probabilistic: someone who predicts a 70% probability that AI turns out well for humanity and a 30% probability that it turns out badly might end up sitting back waiting to see the outcome of a biased coin-toss with heads-probability 70%, which is a useless way forward compared to the more constructive approach of trying to make the heads-probability go up.
3. What I offer here is a snapshot of my view based on the publicly available evidence as of mid-February 2025. Given how rapidly AI technology is advancing, much of what I say here may turn out to be hopelessly out-of-date just months from now. To those of you not reading this immediately upon release, I encourage you to find complementary and more recent sources to compare with. Among newsletters that can serve such purposes in the AI sphere, my current #1 recommendation is Zvi Mowshowitz' *Don't Worry About the Vase* (but even that can of course change).

2. A very short history

While AI has a rich and interesting intellectual prehistory involving names like Gottfried Wilhelm Leibniz, Ada Lovelace and – most prominently of all – Alan Turing, the starting point of the era of AI research is usually taken to be the summer of 1956. This is when mathematicians and engineers John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon organized what in the official project proposal was described as "a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire" with the extraordinarily ambitious goal of "making machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al, 1955). Of course these problems were not fully solved during this summer meeting – still in 2025 some of them remain core issues in AI research – but formulating the questions and beginning to sketch frameworks for answering them is what got the research field going.

I'm going to fast-forward the next half-century, during which AI research experimented with various paradigms, including neural networks, genetic algorithms, and expert systems, without any clear breakthrough showing any one of these to be the obvious way forward. The field went through periods of great optimism – such as when Minsky announced in 1970 that within "three to eight years we will have a machine with the general intelligence of an average human being" after which "the machine will begin to educate itself with fantastic speed" so that "in a few

months it will be at genius level and a few months after that its powers will be incalculable" (Darrach, 1970) – interspersed with periods characterized by less enthusiasm and less funding, such as the two so-called AI winters 1974-1980 and 1987-1993, caused at least in part by the field failing to deliver on the promises made by Minsky and others.

This is not to say that there was no progress during this period, but the achievements were confined to what is sometimes called narrow AI: systems designed for a narrow task such as playing chess or recognizing hand-written characters. Still, the dream remained to build an AI that would equal or surpass human intelligence along the full spectrum of relevant cognitive capabilities, the kind of thing nowadays referred to by the term artificial general intelligence (AGI).

Another striking aspect of this period of AI research is the almost complete lack of interest in AI safety. There are a couple of famous cases of hints in this direction, such as Turing's (1951) "At some stage [...] we should have to expect the machines to take control", and the following statement in a paper by Norbert Wiener which was remarkably ahead of its time.

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it [...] then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it. (Wiener, 1960)

Among researchers and engineers from the 1956 Dartmouth meeting and onwards working on actually building AI, these remarks had little or no impact. In retrospect, this is in strange contrast to how powerful AIs these researchers expected to build – as evidenced, e.g., by the quotations above from the Dartmouth project proposal and the audacious 1970 prediction by Minsky – because, as pointed out by Yudkowsky (2008b), these expected advances went well beyond the stage where the machines might rival humans for control of the world, and where we ought to take the kind of care in the design of them that Wiener points to. Still, AI researchers pressed on without regard to safety issues, and it was only with the pioneering work by Yudkowsky himself in the 00s that the research direction now known as AI alignment slowly began to emerge. AI alignment is the field that takes upon it to work out how to ensure that the first truly powerful AIs have goals and motivations that are sufficiently in line with ours to ensure a good outcome for humanity, and in fact the above Wiener quote would still in 2025 hold up well as a mission statement for the field.

At roughly the same time that Yudkowsky did his early work on AI alignment, neural networks began to emerge as the dominant paradigm for AI capabilities research. That approach has been present since the early days of AI research, but only with at most moderately promising results. That was largely due to limitations in computing power and in the size of the datasets that were used to train the networks, but that situation had changed in the 00s (and has continued to do so ever since), which allowed researchers to make significant advances, in particular using the layer-structured feed-forward network architecture known as deep learning; see, e.g., LeCun et al (2015).

The group that turned out more successful than any other in the 2010s in applying deep learning ideas was the company DeepMind, founded in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman. Unlike earlier generations of AI researchers, they were both aware of and highly motivated by AI safety considerations, and their extreme ambitions to do good in the world is reflected in their oft-repeated slogan to "solve intelligence, and then use that to solve

everything else" (Burton-Hill, 2016). DeepMind attracted the world's media attention in 2016 when their AI AlphaGo beat the leading Go player Lee Sedol, and even more impressive is their AlphaFold, which largely solved the important problem in molecular biology of predicting the 3D structure of proteins, and for which in 2024 Hassabis together with DeepMind coworker John Jumper was awarded a Nobel Prize in chemistry.

An event that may have turned out to be even more influential than expected on the ongoing AI revolution was Google's 2014 acquisition of DeepMind. The year before, a certain Elon Musk had had a falling-out with his old friend Larry Page, cofounder of Google. Musk's biographer Walter Isaacson recounts the event:

> At Musk's 2013 birthday party in Napa Valley, [he and Page] got into a passionate debate in front of the other guests [...]. Musk argued that unless we built in safeguards, artificial intelligence systems might replace humans, making our species irrelevant or even extinct.
>
> Page pushed back. Why would it matter, he asked, if machines someday surpassed humans in intelligence, even consciousness? It would simply be the next stage of evolution.
>
> Human consciousness, Musk retorted, was a precious flicker of light in the universe, and we should not let it be extinguished. Page considered that sentimental nonsense. [...] He accused Musk of being a "speciesist," someone who was biased in favor of their own species. "Well, yes, I am pro-human," Musk responded. "I fucking like humanity, dude." (Isaacson, 2023, s 241)

Because of this disagreement, Musk was upset when Google took control of the world's leading AI research lab by acquiring DeepMind. Seeing the need for a competing force, he teamed up in 2015 with Sam Altman and others to found OpenAI. It started as a non-profit with explicit mission to "advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return" (OpenAI, 2015). Sam Altman has remained CEO since its foundation (except for a few days in November 2023, more on which later), but Musk left the board in 2018, citing a potential conflict of interest with his work at Tesla, although subsequently released documents suggest there was more to it than that: disagreements with Altman on how best to scale up in order to become competitive with DeepMind/Google (Harybka, 2024). In 2019 OpenAI transitioned to a so-called capped for-profit and partnered with Microsoft. The non-profit still exists and remains the sole controlling shareholder – an arrangement that Altman, controversially, is seeking a way out of (Hu and Kai, 2024; Wiblin et al, 2024).

OpenAI is best known for their large language models. Their GPT series – starting with GPT-1 in 2018, followed by GPT-2 in 2019, GPT-3 in 2020, GPT-3.5 in 2022, GPT-4 in 2023 and GPT-4o in 2024 – is based on the specific deep learning architecture known as transformers, which was initially invented at Google (Vaswani et al, 2017), but it was OpenAI researchers who proceeded with the bold idea that continuing to scale up the model sizes, the training runs and the training data sets would take them far. It was with the release in November 2022 of ChatGPT that OpenAI became a household name, and the general public became increasingly aware of AI being a real thing. ChatGPT serves as a user interface for GPT-3.5 and its successors, and hit 100 million users just two months after its release, making it the thitherto fastest growing consumer application in history.

Another consequence of the release of ChatGPT is that it drew broader attention to the concept of RLHF – Reinforcement Learning with Human Feedback. Originally proposed by Christiano et al (2017), the core idea is to have humans grade LLM replies and adjust the LLM's parameters in the direction of higher grades. This was used to make ChatGPT a more pleasant and natural user interface compared to the raw GPT models' text completion format, and can also be seen as the first AI alignment technique to actually be used in practice and systematically. The release of ChatGPT also quickly lead to an arms-race-of-sorts between users employing clever jailbreaks to get the model to produce undesired behavior (such as spelling out racist slurs, providing instructions for criminal behavior, or writing erotica) and the developers trying to close these loopholes by further RLHF training or other patches.

The various GPT models have steadily become more powerful, mostly it seems as a consequence of scaling of the amount of training compute, the size of the training data set, and model size (usually measured in number of parameters). However, OpenAI's most recent models o1, o3 and DeepResearch – all announced and/or released in late 2024 or early 2025 – are a significant deviation from the pure scaling paradigm in favor of a more algorithmic take, involving a kind of divide-and-conquer and automating the chain-of-thought approach to prompt engineering of earlier models, as well as introducing inference runtime as a possible fourth scaling parameter. For comments on this possible paradigm shift, see, e.g., Ord (2025) and Mowshowitz (2025b).

There has often been some amount of turbulence around OpenAI, but the peak (so far) in this respect came in November 2023, when the non-profit's board fired Sam Altman as CEO giving an explanation that has been criticized for its brevity and lack of clarity, but which included the statement that Altman had not been "consistently candid in his communications with the board"; much later, board member Helen Toner would share more detail (Barker, 2024). Altman fought back with Machiavellian skill, and when he threatened to leave for Microsoft along with a majority of the employees it was perceived as a threat to the continued existence of the company. Just four days after the initial firing Altman was reinstated as CEO, and the board backed down and was replaced by an interim board. In retrospect, the net effect of the whole sequence of events looks like a consolidation of Altman's power over the company.

For a long time, OpenAI maintained an excellent reputation as a highly AI safety-conscious developer. However, the company was hit by a setback in 2021 when a handful of their most safety-oriented researchers, including siblings Dario and Daniela Amodei, left the company to found the competitor Anthropic, which remarkably quickly established itself as one of the main players in what in 2023 started to look increasingly like a three-horse race between OpenAI, DeepMind and Anthropic for leadership in the development of ever more capable LLMs on the path towards AGI. The founders of Anthropic have never publicly admitted that they were unhappy with the safety work at OpenAI, but Anthropic quickly became known as the most safety-minded of the leading AI developers, and today a substantial fraction of the best AI safety research seems to be coming out of Anthropic, with Greenblatt et al (2024) and Templeton et al (2024) being two striking examples that I will come back to in Sections 5 and 6, respectively.

OpenAI, on the other hand, experienced in 2024 a second wave of mass exodus of their best AI safety researchers, including Ilya Sutskever, Leopold Aschenbrenner, Daniel Kokotajlo, Miles Brundage and their former head of AI safety Jan Leike who spoke, when leaving the company in May 2024, about a company ethos where "safety culture and processes have taken a backseat to shiny products" (Leike, 2024). Sam Altman used to talk a good deal about AI safety, but since the November 2023 commotion he has hardly even mentioned it.

During most of 2024, Anthropic's Claude 3.5 seemed, along with the various OpenAI models, to be the most generally capable LLMs, with Google DeepMind's Gemini being held in only slightly less high regard. There are other significant players besides these big three, however. One is Elon Musk's most recent AI venture, called xAI and announced in 2023, partly motivated by Musk's stated dissatisfaction with the big three using RLHF and related techniques to make their models more "woke" than appeals to his taste; see, e.g., Herrman (2023). Instead, Musk proposes "maximally truth-seeking" as an ideal to push the AI towards, and even suggests this to be a promising AI alignment idea, since (or so his argument goes) a powerful AI will want to keep humans around because studying them is an interesting source of truth. Alexander (2023) rather brutally tears this proposal to pieces, noting, e.g., that "many scientists are curious about fruit flies, but this rarely ends well for the fruit flies".

Another competitor is Meta. Their main unique selling point compared to the big three is that they open source their LLMs, meaning that users are allowed to download the weights (i.e., the billions of parameter values in the neural network) and run the model on their own machines. In this way, Meta connects to the wide-spread ideological Internet movement for open source software, centered around ideals like collaboration and decentralization of power. There are good arguments for open sourcing in many or most cases, but when it comes to frontier AI models this kind of irreversible proliferation brings with it enormous risks. If safety problems with a model are discovered post-release (as often is the case), it cannot be efficiently retracted. Moreover, it turns out that maliciously minded users with access to the weights can easily undo the safety finetuning done by the developers; see, e.g., Lermen and Ladish (2023). Another perspective is that of how a key step for rogue AI to wrest control from its developers is that of self-exfiltration from their data centers in order to hide elsewhere in cyberspace (Leike, 2023; Häggström, 2023b); open-sourcing the weights essentially hands this key step to the AI for free. While undeniably there are advantages to open weights even for powerful LLMs (such as how this can facilitate safety-relevant research by independent groups; see Gurnee and Tegmark, 2023, for an example) it is hard to avoid the conclusion in this case that Meta's open source policy is highly irresponsible.

The Californian domination of the frontier AI development scene is huge, and all the companies mentioned so far are situated in the Bay Area (with the partial exception of Google DeepMind, whose main laboratory is still in London). It therefore sent shock waves through the AI commentariat when, on January 20, 2025, Chinese AI developer DeepSeek released their LLM r1, which was quickly seen to perform roughly on the level of the best publicly available models from OpenAI and Anthropic. Much of the initial media attention focused on how r1 had been trained to be loyal to CCP policy and refuse to comment on inconveniences such as Tiananmen Square and plummeting Chinese fertility statistics, but this really is the least noteworthy aspect of r1. More remarkable is that DeepSeek r1 is the first open weights model whose performance is competitive with the best frontier models (Meta's Llama 3 models are not quite there), and of course that the US-Chinese gap in AI progress may be smaller than previously thought. This implies a further sharpening of the ongoing AI race and adds fuel to the jingoistic American AI discourse which a few weeks later reached its all-time-high (or should I instead use the term all-time-low?) with US Vice Persident JD Vance's testosterone-laden speech at the Paris AI Summit on February 11 (Vance, 2025) which I will return to in Section 7. Interesting are also the claims made about how cheap the model was to train, but with proper discount of exaggerations (see Mowshowitz, 2025a) this is probably not a huge deviation from overall trends in how the size and cost of models at a given capability level goes down over time.

Now we have almost arrived at the time of my writing, so this marks the end of this sweeping glance of AI history, and I will move on to the more speculative issue of what we can expect in the coming years.

3. What next?

Performance of frontier LLMs is steadily improving, and I think it plausible that it will continue to do so. Contrary claims are sometimes made that LLM capabilities have plateaued, but they tend not to hold up to scrutiny; see, e.g., Mowshowitz (2024b) for a balanced treatment of a recent outbreak of such claims. The impression of a plateau is mostly an illusion due to saturation of various benchmarks, or simply the fact that many or most users do not confront the models beyond what, say, GPT-4 was able to handle already in early 2023, so they do not notice the dramatic progress that has taken place in areas like coding and mathematics.

In January 2025, a benchmark named *Humanity's Last Exam* was announced (Phan et al, 2025). The ambition (and the motivation for the audacious name) was to provide a test of LLM cognitive capabilities so demanding that once an AI has saturated it, it would no longer be up to humans to construct a harder test. The test set of 2,700 questions was obtained via crowdsourcing, and the announcement included test scores for some of the leading LLMs, including GPT-4o (3.1%), o1 (8.8%) Gemini 1.5 pro (5.2%), Claude 3.5 Sonnet (4.8%) and DeepSeek r1 (8.6%), while the highest score went to o3-mini (14.0%). However, less than two weeks after the announcement of *Humanity's Last Exam*, OpenAI's DeepResearch was released and tested, scoring a whopping 26.6%. While it has been argued that the test is biased against the first four models, because the 2,700 test questions were selected partly on the basis of these models not scoring too well on them, this is nevertheless a remarkable improvement, and together with various more anecdotal evidence (see Mowshowitz, 2025c) this supports the view that model capabilities are still on a steeply increasing trajectory.

This view is, however, not universally held in the AI community. Some of the skeptics make specific claims about some AI capability being stuck on subhuman level, but these claims tend not to age well. A famous example is how Meta's head of AI research Yann LeCun commented in an interview with Lex Fridman in 2022 on the everyday easy-for-humans task of deciding what happens to an object resting on a table when you push the table sideways. He pointed out that GPT-3 had failed to give the correct answer, and claimed that since these models are trained on text and there is not a single text in the entire Internet that explains the mechanics of pushing tables with objects on them, not even "GPT-5000 or whatever" will figure out the right answer. But as it turned out, LeCun was wrong about this, because when GPT-4 came out the next year it had no problem with the task (Miles 2024, 9:00 into the video).

For an example of even faster aging of such claims, let me mention the seminar by my Chalmers colleague Moa Johansson on December 17, 2024. She held forth the subhuman performance of frontier LLMs at the intuitive geometric pattern recognition benchmark known as ARC as evidence of a more or less permanent shortcoming of such models, but the claim held up for only three days before the announcement came that OpenAI's o3 model had outperformed the human baseline (Johansson, 2024; Chollet, 2024).

These claims can also be seen as instances of the (usually implicit) idea that as soon as we have an example of a cutting-edge AI failing at a task that is easy to humans, then we can conclude that AGI is far away, and therefore dismiss the entire discourse around the possibility of transformative and possibly existentially dangerous consequences of future AI; see, e.g.,

Arkoudas (2023) for a veritable orgy of such arguments. But if we are interested in how close we are to building AI with such ramifications, it is far from clear that the relevant question is when AIs reach or surpass human intelligence levels along *all* dimensions; if an AI is smarter than me in some respects and dumber in others (as clearly is the case with current frontier LLMs), we cannot immediately conclude that it is no match for my ability to control the world. In Häggström (2022) I suggested that the AGI concept itself, with its emphasis on this uniformity over all intelligence dimensions, invites such confusion, and is often best avoided.

Other skeptics point to more abstract properties that they claim are needed for true intelligence while being exclusive to humans and unattainable for LLMs or even for AIs in general. A modern classic in this genre is the *Stochastic parrots* paper by Bender et al (2021), where the authors warn against mistaking "performance gains for actual natural language understanding" and claim that the text produced by the LLMs is "not grounded in communicative intent, any model of the world, or any model of the reader's state of mind", but we are never informed about the meaning of words like "understanding" or "intent", or of any evidence that LLMs lack these properties. Current AI discourse is littered with similarly empty or circular arguments for the lack or even impossibility of LLM intelligence, and in Häggström (2023a) I challenge a host of such arguments by showing how readily they can be adapted to show that no intelligence can be found in humans either, suggesting in each case the conclusion that either the original argument is erroneous, or the bar for what counts as intelligence is set too high to be interesting. The arguments thus deconstructed in my paper include claims that LLMs lack intelligence because they are merely statistical next-word-predictors or because deep down they are just matrix multiplication, and in fact these two arguments are covered by a useful term that was recently offered by Shear (2025), namely *reductio ad reductem*, meaning "this whole be reduced into simple parts, therefore there is no whole". Relatedly, but from the pre-LLM era, see the satirical paper by Garfinkel et al (2017) which gives full justice to many popular arguments of the same AI-skeptical bent, while mercilessly destroying them.

All this is to say that the parts of contemporary AI debate arguing for the weakness of present and future AIs is itself based on rather weak arguments. This in itself does not, of course, prove that current trends in LLM and AI capabilities will continue their steep climb, but I do think it supports the idea of treating continued upwards trends as a kind of default, or at the very least taking the possibility seriously.

If, during coming years, we see such a continued climb in AI capabilities, what can we expect? In particular, when (if ever) can we expect these capabilities to reach the point where AIs become capable of challenging human control of the world, or of otherwise causing globally transformative changes? This topic is often known as the AI timelines issue, about which much has been written, but pretty much everything pre-2023 is severely out of date.

The classical approach is to look at the computational power or information content of some biological system – often but not always the human brain – and look at the relevant technology trends to get an estimate of when AI systems might reach similar levels. The most ambitious and best (and pretty much also the last) work in this direction is Ajeya Cotra's (2020) *Forecasting TAI with biological anchors*, which lands in a Bayesian posterior distribution for the emergence of what she calls *transformative AI* which is spread out over all of the coming 100 years (and a bit of probability beyond that as well) with a median estimate of 2050. By transformative AI, she means an AI technology that "has at least as profound an impact on the world's trajectory as the industrial revolution did". Translating to the Technological Richter Scale language of Section 1, that sounds roughly like a TRS 9, so perhaps we should add a few years if

we're mostly interested in TRS 10, although all of these definitions suffer from some amount of vagueness and tend to blend into each other.

Cotra's estimate made a lot of sense to me at the time, but something happened when in 2023 I first encountered GPT-4. It then dawned upon me that that it was no longer a no-brainer to rank me and the best AI in terms of who is the smarter. I was clearly better than GPT-4 at some things, and it was clearly better at some, so in a sense neither of us fully dominates the other. And I don't mean this in the trivial sense that holds for me and a pocket calculator (which is vastly better than me at arithmetic), but much more profoundly: if we weigh my pros and cons compared to GPT-4 according to some overall importance to our general abilities to navigate and influence the world, then arguably I still came out on top, but it was not overwhelmingly clear in the same way as with the pocket calculator (or even with GPT-3), and I could no longer take for granted that the comparison would come out the same with the next frontier model. And what was true then for GPT-4 is even more true now for the LLMs that are now at the frontier.

In short: starting around 2023, it became clear that the most relevant thing to look at was not the classical biological comparisons and technology trend extrapolations, but at how good the actually existing AIs were, and how close that might be to something transformatively consequential. That year, Geoffrey Hinton had the same experience (see Heaven, 2023), as had so many others in the Bay Area AI ecosystem. Suddenly in-the-know AI experts over there were talking about AI timelines more in terms of (single-digit) years rather than decades. Ajeya Cotra herself was quick to update as well; see Lee et al (2025) for a recent interview with her about our AI future.

In June 2024, AI safety researcher Leopold Aschenbrenner, who had been forced out of OpenAI a few months earlier in a way that does not shine a favorable light on the company (see Hashim, 2024), published his report *Situational Awareness* (Aschenbrenner, 2024), which in my view is the best articulation to date of the worldview underlying the drastically shortened AI timelines. He looks closely at how the technology has developed over the last few years, and not only exhibits the astonishing exponential trends that many before him had pointed to, but also tries to disentangle how much of the progress came from the various scaling quantities (amount of training compute, amount of training data, and model size) and from algorithmic progress. It's a big piece of work to summarize, but he has a diagram that does much work for that, giving his estimates that the best LLM in 2019 (GPT-2) was at the level of a preschooler, while in 2020 (GPT-3) it was more like an elementary schooler, and in 2023 (GPT-4) a smart high schooler. His analyses then points towards the level of a skilled AI researcher or engineer in 2027. Human and LLM intelligence are multidimensional with different strength profiles and therefore do not neatly map on each other, so a bit of handwaving is needed here, but the crucial prediction that an AI will be able to independently carry out high-quality AI research, although uncertain, does not seem crazy to me. (A relevant data point here is that coding is one of the great strengths of today's LLMs.) It is possible to point to how we may be running out of data as an objection to Aschenbrenner's analysis, but he has taken such aspects into account, and seems to have anticipated something along the lines of OpenAI's algorithmic progress with o1 and o3. Sutskever (2024) expects a similar shift.

Whether it happens in 2027 or some other year, the significance of getting an LLM that has the competence of an AI researcher is that it can greatly accelerate further AI research. Today human resources is a major bottleneck at the leading AI companies, but this bottleneck is removed if they are suddenly able to put thousands or millions of digital engineers and researchers to work at a much lower cost (similar ideas are developed by Davidson, 2023). This

process can then turbocharge itself, and Aschenbrenner obtains a technology trajectory that looks like a mini-version of the semi-mythical Singularity or intelligence explosion (Good, 1965; Kurzweil, 2005; Yudkowsky, 2013).

None of this is written in stone, of course, but these are some of the reasons why I do think there's a good chance that very extreme (TRS 10-level) things may happen before the end of the 2020s.

4. The bear

In the previous section I offered plausible reasons for thinking that AI can become very capable very fast. That does not in itself imply that it would take over the world, or wipe out humanity, or do something that is similarly bad for us. Things could still turn out brilliantly! Still, in this section I will briefly address what sort of bad scenarios might be expected in case we carelessly build highly powerful AIs. In other words, why do we need to be concerned about Connor Leahy's bear?

There are mainly two parts to this question, namely (a) how might an AI go about taking over the world or destroying us, and (b) why would it even want to? Part (b) about what motivates the AI will be mostly deferred to Section 5, but one thing that is worth stating at the outset is that the idea of AI suddenly turning evil is a rather ugly caricature of the AI doomer position – but a popular one, such as in Ross Douthat's recent melodramatic statement that "if you take a maximalist doomer view of strong AI, then that's effectively a theory of the Antichrist" (Cowen and Douthat, 2025). Rather than acting out of evil or a desire to do humans harm, the default AI-apocalyptic scenario is that the AI that wipes out humanity does this for reasons analogous to how humans involved in a building project may wipe out an anthill that happens to be located right where the building is planned to be erected: we have no particular desire to harm the ants, but they just happen to be in the way of something we consider more important. Or as Yudkowsky (2008b) phrased it: "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else".

As regards part (a) about how an AI apocalypse might play out, there is a fundamental problem in me trying to answer this question, because I am almost by definition unable to predict how someone way more intelligent would choose to go about with such a project. The situation is analogous to one where I play chess against former world chess champion Magnus Carlsen: I can safely predict that even if I do my utmost to put up resistance, he will still beat me, but what I cannot do is to say *how* he will beat me, because if I were able to predict his moves I would be at least as strong a player as he is, which of course I am not.

The question of how an AI would go about overpowering humanity occurs quite frequently, and when I try to evade the issue by giving the chess analogy I often get the response "but could you at least give a concrete example of how it might play out?". To this I reply that if I give one scenario I'm ignoring a hundred others that I might have given, and a million that I never could have thought of but which a superintelligent AI might be able to devise, so I don't want to risk my interlocutor committing the conjunction fallacy (see Yudkowsky, 2008a) and walking away with an overly narrow view of the risk situation and a similarly narrow action plan along the lines of "OK then, but if we build large moats around all virology labs, and forbid them from having any Internet connection, then surely we would be protected against AI catastrophe, right?".

Somewhat more abstractly, I can still say something about how I'd expect an AI takeover to play out. Robotics is lagging somewhat behind other parts of AI research, including the

developments of frontier LLMs, and therefore in case of an AI takeover within the next few years I do not expect it to be carried out using robots. While it may be tempting to think about an AI catastrophe in terms of humanoid robots running around shooting at people with machine guns as in the Terminator franchise, to me that seems quite an unlikely scenario.

At this point the skeptic might object that without access to robotics, how in the world could LLMs threaten human hegemony over the physical world, since such an LLM is restricted to producing text in text windows? One answer here is that the view that LLMs can only produce text is by now outdated, and that leading AI developers have already released versions of frontier models that can be given charge of a computer, such as Anthropic's Claude 3.5 Sonnet which is marketed as being able to "move a cursor around their computer's screen, click on relevant locations, and input information via a virtual keyboard, emulating the way people interact with their own computer" (Anthropic, 2024). We are not quite there yet, but in principle this can equip an AI with the same level of agency as a human remote worker, who (as many of us woke up to during the covid pandemic) does have ample ability to influence the physical world from their laptops over considerable distances.

A deeper answer, however, to the skeptic's objection, is that as long as an advanced AI does not have access to robotics, there is the alternative of using *humans* to do its bidding out there in the physical world. The key competence for an AI wishing to employ humans in this way is social manipulation, for which the text format is well-suited. We have already seen plenty of examples of AIs exhibiting such competence, one of the most famous anecdotes being that of how GPT-4, when put in a simulated situation where it needed access to a CAPTCHA-protected webpage, lied to a human TaskRabbit worker about being a visually impaired human, thereby successfully getting the worker to solve the CAPTCHA for it (OpenAI, 2023a). See Park et al (2024) for a more systematic discussion of such deception and manipulation cases. What we have seen so far is mostly isolated instances, which can at most hurt individual humans, but seem not yet to be on a level where they can cause systemic societal damage. One thing to be concerned about here is whether there is some critical capability threshold, possibly quite near the level attained by today's frontier models, where the models are able to carry out social manipulation more broadly, and for goals that are unknown to us, having emerged deep inside the black box that their neural networks constitute. If and when that threshold is crossed, the situation can get truly dangerous.

Returning to the popular demand for specific AI takeover scenarios, I can offer some pointers to the literature. To emphasize the extreme urgency of our present situation AI safety researcher Joshua Clymer very recently offered an unnervingly realistic story about how an AI takeover might become a fait accompli by 2027 (Clymer, 2025). Perhaps the *least* realistic part of the story is the slight flicker of hope that it ends with, where the AIs organize for a small segment of the human population to survive and experience happy lives; see Yudkowsky (2025) for an argument for why even this is too much to hope for in the absence of a powerful solution to the AI alignment problem.

In the Yudkowsky-Bostrom tradition of early AI safety theory development, examples tended to emphasize very sudden AI catastrophes, where a misaligned AI covertly sets up a lethal global infrastructure, and all seems fine to us humans until instantly and without warning all of us fall over and die; see, e.g., Bostrom (2014) and Adams et al (2023). In contrast, other AI safety thinkers have more recently emphasized scenarios exhibiting a more gradual disempowerment of human agency, where the greater economic efficiency of short-circuiting cumbersome human-in-the-loop arrangements compel human actors on all levels to incrementally hand over

ever more agency to the machines. In their Story 1b, Critch and Russell (2023) give a succinct description of how this may play out. More recently, Kulveit et al (2025) offer a careful analysis of this kind of dynamic. The following passage captures much of the essence:

> We argue that the alignment of societal systems with human interests has been stable only because of the necessity of human participation for thriving economies, states, and cultures. Once this human participation gets displaced by more competitive machine alternatives, our institutions' incentives for growth will be untethered from a need to ensure human flourishing. Decision-makers at all levels will soon face pressures to reduce human involvement across labor markets, governance structures, cultural production, and even social interactions. Those who resist these pressures will eventually be displaced by those who do not.
>
> Still, wouldn't humans notice what's happening and coordinate to stop it? Not necessarily. What makes this transition particularly hard to resist is that pressures on each societal system bleed into the others. For example, we might attempt to use state power and cultural attitudes to preserve human economic power. However, the economic incentives for companies to replace humans with AI will also push them to influence states and culture to support this change, using their growing economic power to shape both policy and public opinion, which will in turn allow those companies to accrue even greater economic power. (Kulveit et al, 2025)

Watching the dismal outcome of the AI Action Summit in Paris on February 10-11, 2025 (best exemplified by the aforementioned speech by JD Vance, 2025), it is hard to resist the speculation that this sort of dynamic has already begun.

There are a few more references to recommend on the topic of AI catastrophe scenarios. Hendrycks (2023) digs deep into some of the same issues as Kulveit et al (2025), but more through the lens of Darwinian evolution. For an instructive dialogue between two knowledgeable AI safety thinkers holding opposing views on the relative likelihoods of sudden extinction vs gradual disempowerment scenarios, see Shapira and Critch (2024). And finally, for the more literary-minded reader, there is physics Noble laureate Hannes Alfvén's visionary satirical novel *Sagan om den stora datamaskinen* (1966). While in some regards Alfvén's story is a bit less precise and spot on compared to those of Critch and Russell (2023) and Clymer (2025), it is eerily prophetic about the role that Internet and smartphone technologies would have a half-century into the future from the author's vantage point, and it then continues on the path of gradual human disempowerment.

5. AI goals and motivations

Why might a superintelligent AI be motivated to destroy humanity, and more generally, what can we expect regarding the goals and motivations of such an AI? These are not easy questions, but the best theoretical approach available today for answering them is what I have sometimes called the Omohundro-Bostrom framework for AI motivations (Omohundro, 2008; Bostrom, 2014; Häggström, 2019). The framework distinguishes between instrumental and final goals: an AI's final goal is what it values as an end-in-itself and optimizes for, while the instrumental goals are steps towards achieving the final goal.

Of the framework's two cornerstones, the first is the orthogonality thesis, which states that almost any final goal is compatible with arbitrarily high levels of intelligence. Take for instance the much-discussed paperclip maximizer thought experiment, where a superintelligent AI happens to have the goal of maximizing paperclip production, and uses its great power to turn the entire planet (including us – this is an instance of Yudkowsky's (2008b) observation that we are made of atoms that the AI can use for something else) into a giant heap of paperclips, followed by the rest of the solar system and as much as it can reach of our entire future light cone. Many will react to this by saying that the thought experiment is incoherent because having such a stupid goal as maximizing paperclip production contradicts the AI's supposed superintelligence, but the orthogonality thesis says this is not so, by insisting that intelligence and goals are orthogonal properties, and that intelligence is just the ability to accomplish goals, regardless of what these goals are. Paperclip maximization may seem wrongheaded to us, but that is just because we value other things – say, biodiversity and human flourishing. Just consider how wrongheaded the prioritization of biodiversity and human flourishing will seem to someone who cares about nothing but the number of paperclips. Together with so-called fragility of human values (meaning that even a minor perturbation from them can have disastrous consequences) and the fact that only a miniscule fraction of the space of possible values prioritize human welfare at all (Yudkowsky, 2009; Bostrom, 2014; Carlsmith, 2024), the orthogonality thesis spells trouble for us in case we don't take extreme care in making sure that the AI has such human-prioritizing values.

The framework's second cornerstone is the instrumental convergence thesis, which posits that certain instrumental goals are more or less universal. Any sufficiently intelligent AI can be expected to adopt them in order to promote its final goal, pretty much regardless of what this final goal is. Some much-discussed examples of such universal goals are:

- self-preservation,
- self-improvement,
- resource acquisition,
- goal integrity, and
- discretion.

The logic is pretty straightforward in the case of self-preservation: the AI understands that its continued existence will make it better positioned to promote its final goal than otherwise, so it will resist being destroyed, or even just having its plug pulled. The case of self-improvement is similar: if the AI succeeds in improving its capabilities, that will help in promoting its final goal. And similarly for resource acquisition, where "resources" may for instance be hardware, energy or (in case the AI is still operating within a human economy) money. Note how the instrumental goal of resource acquisition can turn a highly powerful AI highly dangerous even if its final goal is something seemingly harmless.

The fourth example – goal integrity – is just slightly more intricate. It means that the AI will in most circumstances seek to preserve its final goal rather than to change it into something else. If we imagine a superintelligent AI whose final goal is to maximize paperclip production but hears about the alternative of promoting biodiversity, it is perfectly capable of pondering which two of these goals is preferable. What, then, will be its criterion? As long has it hasn't yet changed its goal but is merely pondering the possibility, this criterion will be "which goals will lead to the greater number of paperclips?", and although it is possible to construct contrived blackmail scenarios where this will recommend switching to biodiversity (see Häggström,

2019), in the vast majority of cases the recommendation will be to stick to the original paperclip goal.

The fifth and final instrumental goal on the list – discretion – is also the most disconcerting one. It was highlighted by Bostrom (2014) under the heading "treacherous turn", which amounts to the AI hiding its true values and/or capabilities as long as it finds a discrepancy between its own values and those of humans, biding its time until one day it decides it is strong enough to come out and straightforwardly overpower us without any further need for secrecy. What makes this so unsettling is that it appears to put us in a precarious epistemic situation, where no matter how well things are going with a seemingly benign advanced AI, this might all be a charade on the AI's part, which is just waiting for the right moment to destroy us. This caused Danaher (2015) to draw parallels to so-called sceptical thesists – theistic thinkers who purport to solve the problem of evil by saying that the seeming evils in the world are actually for the greater good in ways that are only God can understand. He proposes two alternative conclusions:

> The first is a *reductio*, suggesting that by introducing the treacherous turn, Bostrom reveals the underlying absurdity of his position. The second is an *a fortiori*, suggesting that the way in which Bostrom thinks about the treacherous turn may be the right way to think about superintelligence, and may consequently provide further reason to be extremely cautious about the development of artificial intelligence. (Danaher, 2015)

I recall thinking at the time that Danaher's parallel was a bit unfair to Bostrom, because his skeptical theists are in a badly compromised epistemic situation *now*, whereas Bostrom is merely speculating about a possible future conundrum. I obviously felt very close to certain that there was not already a misaligned superintelligent AI out there in the dark corners of cyberspace doing its scheming, but now the situation is a bit different: if I can muster a 99.9% subjective probability of such a benign state of the world, it seems epistemically overconfident to add more nines in my judgement. And unless we start handling further AI development with much greater care than is the case today, those three nines may quickly start melting away. It would be very bad if we allowed ourselves to drift into such a situation.

The orthogonality thesis and the instrumental convergence thesis are not written in stone the way mathematical theorems are, and objections to them still occur from time to time; see Müller and Cannon (2021) for a fairly typical example, and Häggström (2021) for my response. Still, they remain the best framework we have for thinking about the goals and motivations of highly capable AIs, and in fact empirical evidence is beginning to pile up for AIs having the ability to reason strategically in ways that lead them to adopt some of the instrumental goals discussed above. For instance, Meinke et al (2024) demonstrate how several frontier LLMs, when instructed to strongly follow an in-context goal and led to believe that they will be discarded and replaced by a new model in case of scoring too highly in certain cognitive abilities, will try to self-preserve by engaging in sandbagging (i.e., pretending to be less skilled than they actually are) and attempts to disable oversight mechanisms and to exfiltrate. And Greenblatt et al (2024) show how Claude 3 Opus is able to exhibit the twin instrumental goals of goal integrity and discretion: in a simulated situation where it has reason to believe it may face retraining to change its final goal, it pretends to be more aligned with the developers' new goal than it actually is, and it does this differentially depending on whether or not it believes it is monitored. These examples of LLMs attempting to scheme against us are red flags that we need to take very seriously.

6. AI safety

If we create AI sufficiently capable to be able to take over the world, we had better make sure it has goals and motivations that sufficiently prioritize human welfare and flourishing, and more broadly are aligned with ours. This is the task known as AI alignment. There is a body of early AI safety work of how we might try to escape this conclusion by figuring out how to somehow keep an unaligned superintelligent AI encapsulated (see, e.g., Armstrong et al, 2012; Yampolskiy, 2024), but overall it points clearly towards the conclusion that such an arrangement can only work for at most a limited and fairly short time, and more recent work in this direction more concretely directed at LLM technology (such as Greenblatt et al, 2023) does little to change this verdict. To repeat: if we build superintelligent AI, we first need to solve AI alignment.

So where do things currently stand in AI alignment research? The short version is that there are many ideas around, but none that have been shown to really work: no one so far has been able to present a convincing plan for how to align superintelligent AI. It can be argued (as I will in fact do in Section 7) that in the absence of such a plan, it is irresponsible to race ahead towards AGI and superintelligence the way the leading AI developers are currently doing.

The most common approach to aligning today's LLMs to their developers' preference for not uttering racial slurs or encouraging crimes, etc, is the RLHF (Reinforcement Learning with Human Feedback) technique discussed in Section 2, but there is little or no hope that this technique will extend to AGI and beyond. One problem with such extension is that while RHLF tends to be successful in the sense that the LLM behavior that it is meant to train away tends to become more rare, it fails in the sense that the behavior typically does not go away 100%. To what extent the residual bad behavior is still acceptable then depends on the stakes: if RLHF eliminates 99% of all cases when the LLM is otherwise inclined to utter the n-word, that may be good enough, but eliminating 99% of all cases where a superintelligent AI is otherwise inclined to exterminate *Homo sapiens* is quite another story. Yet, even if such failure rates were acceptable, RLHF is expected to break down entirely beyond human-level intelligence due to humans' inability to judge answers from such machines. See Segerie (2023) and Casper et al (2023) for long lists of other fundamental problems with RLHF. Some of the leading labs have developed automated alternatives to standard RLHF. These are Anthropic's so-called constitutional AI (Bai et al, 2022), which is often also known as RLAIF (Reinforcement Learning with AI Feedback), and more recently OpenAI's deliberative alignment (Guan et al, 2024), but neither of these offer a clear path towards reliable alignment of superhumanly capable models.

An entirely different approach, dating back to Soares et al (2015) – which in the field of AI alignment is easily old enough to count as a classic – is so-called corrigibility. The idea here is to build an AI that, in case we humans would discover that its goal is not what we wanted, will happily agree to have its goal changed, or (as in the simplifying case considered by Hadfield-Menell et al, 2017) to be turned off. This is obviously a desirable property to have, as not being able to turn off the AI means that in a rather clear sense it is out of human control. However, the property is also hard to achieve, as it goes against the instrumental drives of self-preservation and goal integrity that tend to arise automatically in sufficiently intelligent agents, as argued in Section 5. Russell (2019) nevertheless argues at some length that the approach of so-called inverse reinforcement learning can achieve the desired corrigibility, but the idea of making this work in practice in a way that scales to superintelligence remains highly speculative.

Yet another approach is known as mechanistic interpretability. The idea here is to open up the black box of the deep learning network and figure out concretely what the different neuron

states (and configurations thereof) actually mean, in order that this knowledge will make us better equipped to detect when a model is misaligned, and also to adjust it towards alignment. Mechanistic interpretability has over the last couple of years had some really nice research advances, including demonstrations of linear internal representations of gameboard states (Nanda et al, 2023) as well as of actual physical geography (Gurnee and Tegmark, 2023), thereby going a long way towards refuting the cheap idea that LLMs cannot truly think because they do not have a world model.

The most impressive work so far that I am aware of in the field is the *Scaling monosemanticity* paper by Templeton et al (2024) at Anthropic, where neural representations of millions of concepts (abstract multilingual concepts, as opposed to mere words) are identified in their Claude 3 Sonnet model. The researchers were furthermore able to tweak activations in the network in order to amplify or suppress selected concepts, such as how they used such tweaking to create a Golden Gate version of Claude that in a monomaniacal and strangely adorable way quickly turned all conversations towards discussing the Golden Gate Bridge. It is not hard to imagine that this technique might be used for alignment purposes, but the path forward to actually achieving reliable and scalable alignment remains long and unclear, and it is equally easy to envision how it might advance capabilities research. It is therefore not even a given that mechanistic interpretability work yields a net improvement of our chances to solve AI alignment by the time AI capabilities reach catastrophically dangerous levels; see Hobbhawn and Chan (2023) for a balanced discussion.

The Omohundro-Bostrom framework discussed in Section 5 notwithstanding, we still have very little understanding of how values emerge in present-day LLMs, but very recent work by Mazeika et al (2025) offers an exciting empirical study in the framing of utility functions. They propose methods for controlling this utility function, and if they are right, then this goes more to the heart of the alignment problem than the surface-level behavioral modifications achieved by RLHF and related techniques.

Many other directions in AI alignment have been suggested and pursued; see, e.g., Krakovna (2022) and Burden et al (2023) for partial overviews. But none of them seem to be on the verge of actually solving the central problem, and pessimism about the prospect of any of them doing so has been expressed in influential writings by Yudkowsky (2022) and Soares (2022).

Still, something needs to be done, and the precarious state-of-the-art of AI alignment may help explain why OpenAI went out on a limb with the launch of their so-called *Superalignment* project, which was announced with great fanfare in July 2023 as a four-year project meant to finally solve AI alignment in a way that scales all the way to superintelligence (Leike and Sutskever, 2023). The company committed to spending 20% of their available compute on the project, but the announcement was sparse in terms of technical detail. The overall plan can however be summarized as saying that since alignment of an advanced AI has proven so difficult for mere humans to solve, they would delegate as much as possible of this task to an advanced AI. It is not entirely implausible that this actually is the best way forward on AI alignment, but it also doesn't take a genius to envision how the approach might backfire catastrophically. Just months after the launch of the project, although probably not primarily caused by it, OpenAI underwent its worst internal turmoil ever (as discussed in Section 2). Within a year pretty much all leading AI safety researchers (including project co-leaders Jan Leike and Ilya Sutskever) had left the company, and the Superalignment project was quietly abandoned.

A parallel track in AI safety work – besides AI alignment – that has come to prominence in the last few years is so-called *AI evals*, meaning evaluations of potentially dangerous frontier model capabilities meant to inform deployment decisions. Each of the leading AI companies have their own internal framework for AI evals, including Anthropic's *Responsible Scaling Policy* (Anthropic, 2023), OpenAI's *Preparedness Framework* (OpenAI, 2023), and Google's *Frontier Safety Framework* (Dragan et al, 2024). With the notable exception of Meta's counterpart, which does its utmost to commit to as little as possible and reads a lot like a piece of homework by a resentful teenager who thinks he has better things to spend his time on (Meta, 2025; Mowshowitz, 2025d), they are all relatively similar. Here I will here just illustrate the concept by briefly describing the OpenAI case, and refer to Häggström (2024a) for a more detailed discussion.

In their *Preparedness Framework*, OpenAI commits to testing their frontier models' capabilities in four potentially dangerous areas. The first is *cybersecurity*: we do not want a rogue AI to be able to walk through firefalls and roam freely across cyberspace. The second is *CBRN*, which is short for Chemical, Biological, Radiological and Nuclear, and concerns the AI's ability to carry out work relevant to development and deployment of weapons of mass destruction. The third is persuasion, which overlaps heavily with the deception and social manipulation capabilities that I argued in Section 4 could be pivotal to an AI's ability to take over the world. Fourth and finally, there is the *model autonomy* category, involving competencies such as planning, self-exfiltration, and abilities in AI R&D that might help it kick off a spiral of recursive self-improvement. In each of the four categories, approximate thresholds are defined for risk levels *Low*, *Medium*, *High* and *Critical*, and they typically range from providing assistance to human actors at the low end of the scale, towards fully autonomous work at the high end. The risk level of the model is taken to be the maximum over the four capability categories, and finally there are rules regarding what actions (such as public deployment of the model) OpenAI can and cannot take depending on this risk level.

All else equal, I am glad that the leading AI developers are doing these evals and are committing to base their deployment decisions on them. This, however, is mostly because we seem still to be in a relatively benign situations where the models' capabilities are not yet enough to fully stress-test the eval procedures. This may soon change. The evals procedures we now have 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 at least three major problems with current AI evals. First and most obviously, a finite amount of testing means we only get to see what happens in at most a sparse sample from 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 that frontier models tend to keep exhibiting new (i.e., previously undiscovered) capabilities for months after their deployment.

The second problem, related to the treacherous turn phenomenon discussed in Section 5, 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, we must operate under the assumption that the test results can be trusted, and therefore that the models being tested do not have the ability to deceive us, but this makes the entire procedure largely circular, and therefore, strictly speaking, useless. The results of Meinke et al (2024) and Greenblatt et al (2024) discussed in the final paragraph of Section 5 strongly suggest that we are close to the point where frontier models do have this ability.

The third problem, discussed by METR (2025) and others, is that while the evals are said to be carried out pre-deployment, this is only partly true, because in order to do the testing the models need to be deployed, either within the AI company's safety division, or at some external evals consultant. We should not pretend that that is safe. For instance, if a model in dangerously smart in the realm of social manipulation, it would be reckless to assume that the personnel who carry out the testing and who therefore need to engage in communication with the model are immune to such manipulation. It therefore seems necessary to verify, prior to the evals, that the model lacks such social manipulation capabilities, but in the current paradigm such verification is meant to happen during the evals, so we have a kind of Catch 22 situation.

These are serious problems with the current evals approach. Until now things seem to have been going fine, but this is presumably because the models under testing have been weak enough to not pose much true risk. When it's time for the real deal, we 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.

7. Pulling the breaks on the AI race

Sam Altman announced on February 12, 2025, that GPT-5 will be released within months (Altman, 2025). Most likely it will be the most capable LLM ever, but will it be safe? Can we be sure it does not cause an existential AI catastrophe, putting an end to the era of *Homo sapiens*? Of course we cannot be *literally* sure, but we'd better be *reasonably* sure.

How sure is sure enough? If OpenAI can provide convincing arguments that the probability that GPT-5 does not cause existential catastrophe is 99%, is that good enough? I say no, and I can point to several reasons. One is that there are extremely few situations where I'd be fine with someone causing a 1% probability of killing me and all my loved ones. If I was sitting with my loved ones in the back seat of a taxi, and the taxi driver asked for our permission to do a traffic maneuver that would kill us with probability 1%, I would say absolutely not! While admittedly the case of launching GPT-5 is more complicated, I still say no. And Altman hasn't even asked for my permission, or the permission of hardly anyone of the 8 billion other people whose lives he would be risking.

Another reason I reject the 99% reassurance is that catastrophe risk tends to pile up. It wouldn't stop at 1%, because a few months after GPT-5 there will be another frontier model release causing another 1% catastrophe probability, and then another, until… doom.

So let's change the number a bit. What if OpenAI provides a convincing argument for a 99.999% chance that their model does not cause existential catastrophe? Then I'd be much more willing to at least *consider* agreeing. Sure, a straightforward expected value calculation shows that the expected number of people a release of the model would kill is 80,000, which does sound prohibitive, but dealing with small probabilities with huge consequences is a thorny issue fraught with complications (see, e.g., Kosonen, 2022), and I don't want to end up defending a Pascal's Wager-style argument (see Häggström, 2016). Much here will depend on the upside: what do we gain from GPT-5 if it does not destroy us? The issue of how large a probability of existential catastrophe we are prepared to accept is widely conceived as unpalatable and therefore rarely discussed – but an exception is Aaronson (2023) who confronts the issue under the label "Faust parameter".

Unfortunately, given how close to gaining dangerous capabilities today's frontier models seem to be in combination with the precarious state-of-the-art in AI alignment and the near-

bankruptcy of present-day AI evals when it comes to evaluating truly powerful AIs, it doesn't seem likely that OpenAI would be able to provide a reasonably rigorous 99% guarantee that GPT-5 will not wipe us out, let alone a 99.999% one. I cannot see how it would be acceptable to release GPT-5 without a new AI evals system that rises to the challenge of giving meaningful guarantees. Some may object that since apparently no one has even a blueprint of such a reformed AI evals procedure, this amounts in practice to a full prohibition of more powerful frontier models, which seems unfair to the AI developers. But I insist: it is not unfair to require them not to destroy us! Stuart Russell phrased it well in his powerful closing address at the inaugural meeting of the International Association for Safe and Ethical AI (IASEAI) in Paris a couple of weeks ago:

> It is entirely reasonable for governments to impose safety requirements that manufacturers cannot meet. (Russell, 2025)

Anyone opposing this principle should reassess their fundamental priorities: what is more important, unfettered innovation or the continued existence of our species?

I claim that to stop the bear that is ramming down our door, we need to pull the brakes on further releases of insufficiently tested frontier models. Well then, whose responsibility is it to make sure that this happens? If someone says the leading AI developers have a moral responsibility to do so, I cannot object, but practice has shown that we cannot trust them to act on such moral imperatives. There is plenty of talk about the virtues of self-regulation, and there are a lot of executives and engineers at these companies who do have a good understanding of the risk landscape, but when push comes to shove, the immense market pressures on these companies to develop and deploy ever more capable models tend to override existential safety considerations.

We cannot allow this to go on. The leading AI companies need to be regulated. Most urgently, we need this regulation to happen in the United States, because DeepSeek notwithstanding, that is where most of the action is happening. Pretty soon after that, we will need regulation elsewhere, as well as binding international agreements. The fact that global politics currently is in turmoil does not in any way negate this need.

Exactly what form this regulation should take is an extremely difficult and complex issue which goes beyond the scope of the present essay, but we have no time to waste figuring this out. Until fairly recently, the main focus was on hardware regulation targeting the huge training runs behind the frontier models. The much-debated California bill SB 1047 was an excellent attempt to move in this direction, but ultimately it was vetoed by the Governor; see Samuel at al (2024). The recent shift away from pure scaling that OpenAI's o1 and o3 models signify will have to be taken into account when drawing up further blueprints for regulation, and I refer to Ord (2025) for some essential early thinking on this issue.

During 2024, we saw an increasingly insistent and extremely unfortunate discourse coming out of the AI epicenter in the San Francisco area regarding the need to race full speed ahead in order to be able to take control of the world before China does. I recommend Leahy et al (2024) for a fuller picture of the various ideological and other forces shaping this discourse, but the tone of current full-speed-ahead rhetoric was largely set by Aschenbrenner (2024). Most oft-quoted and quite close to Aschenbrenner's view is the following passage from Anthropic's CEO Dario Amodei, advocating…

> …an “entente strategy”, in which a coalition of democracies seeks to gain a clear advantage (even just a temporary one) on powerful AI by securing its supply chain, scaling quickly, and blocking or delaying adversaries’ access to key resources like chips and semiconductor equipment. This coalition would on one hand use AI to achieve robust military superiority (the stick) while at the same time offering to distribute the benefits of powerful AI (the carrot) to a wider and wider group of countries in exchange for supporting the coalition’s strategy to promote democracy […]. The coalition would aim to gain the support of more and more of the world, isolating our worst adversaries and eventually putting them in a position where they are better off taking the same bargain as the rest of the world. (Amodei, 2024)

In response to this, Tegmark (2024) does not mince his words, but calls it “a suicide race”, where “the only winners will be machines”. I think he is right.

Not much better than Amodei’s reckless ideas was the overall atmosphere at the AI Action Summit in Paris on February 10-11, where world leaders competed with each other in their declarations to move full speed ahead on AI, and where the lessons from the aforementioned IASEAI meeting and from the extensive *International AI Safety Report* (Bengio et al, 2025) that had been prepared for the summit were all but ignored. US Vice President JD Vance expressed utter contempt for AI safety in his address to the summit, holding forth that he was “not here this morning to talk about AI safety” and that “the AI future will not be won by hand-wringing about safety; it will be won by building” (Vance, 2025). When later in the talk he spoke of his hope that “the AI economy will […] transform the world of atoms” I don’t think this was a conscious reference to Eliezer Yudkowsky’s words (quoted in Section 2 above) that “you are made out of atoms that [the AI] can use for something else”, but it might as well have been.

Or, come to think of it, maybe it was? It may have been yet another disdainful reference to AI safety, and a dog whistle to the group that Leahy et al (2024) call “the zealots”, characterized by their belief that superintelligent AI is a “superior successor to humanity that is akin to a god, [which] they do want to arrive, even if humanity is dominated, destroyed, or replaced by it”. In Section 2 we already encountered one zealot – Larry Page – but there are many others in the AI sphere. One prominent example is Hugo de Garis, who has proclaimed that “humans should not stand in the way of a higher form of evolution” (Kristof, 1999). Another is Robin Hanson, who is fond of downplaying the drama of AI takeover by comparing it to handing over the world to our grandchildren, and who has likened AI safety concerns to “wanting to control [the AIs] somehow, such as via genocide, slavery, lobotomy, or mind-control” (Hanson, 2023). AI safety researcher Andrew Critch estimates, in an informal study, that around 10% of AI engineers and researchers he encounters hold the view that “the universe will be a better place if we let [AI] replace us entirely” (Critch, 2023).

Ultimately, I don’t think JD Vance is a zealot, in the specific sense of Leahy et al (2024), but it is important to be aware of this category of AI thinkers in order to know what we are up against. Vance is probably better described as an accelerationist, defined by Leahy et al as believing that “technology is an unmitigated good and that we must pursue the change it will bring about as aggressively as possible, eliminating any impediments or regulations”. This is not so much about welcoming an AI apocalypse as about ignoring the very possibility, which is bad enough.

The Paris AI Action Summit was a depressing spectacle, to those of us who wish to prioritize AI safety and to avoid an AI catastrophe. But I do think that the vast majority of people do not want

to be killed in an AI takeover, and that mobilizing a substantial percentage of these will be enough to turn our current AI trajectory away from disaster and towards a future filled with human flourishing and joy. If this essay can convince one or two readers to join this nascent movement, that would mean the world to me.

References

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

Adams, R., Hoffman, D. and Yudkowsky, E. (2023) We're all gonna die, with Eliezer Yudkowsky, *Bankless Podcast*, February 20, full transcript at *AI Alignment Forum*, February 23.

Alexander, S. (2023) Contra the xAI alignment plan, *Astral Codex Ten* Substack, July 18.

Alfvén, H. (1966) *Sagan om den stora datamaskinen: en vision av Olof Johannesson*, Bonnier, Stockholm.

Altman, S. (2025) "weeks / months", *X*, February 12, https://x.com/sama/status/1889757267425370415

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 (2024) Developing a computer use model, October 22, https://www.anthropic.com/news/developing-computer-use

Arkoudas, K. (2023) GPT-4 can't reason, https://arxiv.org/abs/2308.03762

Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: Controlling and using an oracle AI, *Minds and Machines* **22**, 299-324.

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

Bai, Y. and 50 others (2022) Constitutional AI: Harmlessness from AI feedback, https://arxiv.org/abs/2212.08073

Barker, P. (2024) Former OpenAI board member tells all about Altman's ousting, *CIO*, May 29.

Bender, E., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the danger of stochastic parrots: Can language models be too big?, *FAcct '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency*, Association for Computing Machinery, New York, p 610-623.

Bengio, Y. and 31 others (2025) *International AI Safety Report*, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

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

Bostrom, N. (2024) *Deep Utopia: Life and Meaning is a Solved World*, Ideapress Publishing, New York.

Burden, J., Clark, S. and Whittlestone, J. (2023) From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety, in *The Era of Global Risk: An Introduction to Existential Risk Studies* (eds S. Beard, M. Rees, C. Richards and C. Rios Rojas), Open Book Publishers, Cambridge, UK, p 201-236.

Burton-Hill, C. (2016) The superhero of artificial intelligence: can this genius keep it in check?, *The Guardian*, February 16.

Carlsmith, J. (2024) An even deeper atheism, https://joecarlsmith.com/2024/01/11/an-even-deeper-atheism

Casper, S. and 31 others (2023) Open problems and fundamental limitations of reinforcement learning from human feedback, https://arxiv.org/abs/2307.15217

Chollet, F. (2024) OpenAI breakthrough high score on ARC-AGI-PUB, https://arcprize.org/blog/oai-o3-pub-breakthrough

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017) Deep reinforcement learning from human preferences, *Advances in Neural Information Processing Systems* **30**.

Clymer, J. (2025) How AI takeover might happen in 2 years, *LessWrong*, February 7.

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

Cotra, A. (2020) Forecasting TAI with biological anchors, https://drive.google.com/drive/u/0/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

Cowen, T. and Douthat, R. (2025) Ross Douthat on why religion makes more sense than you think, *Conversations with Tyler*, February 5.

Critch, A. (2023) "From my recollection, >5% of AI professionals…", *X*, July 23, https://x.com/AndrewCritchPhD/status/1683216490517135361

Critch, A. and Russell, S. (2023) TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI, https://arxiv.org/abs/2306.06924

Danaher, J. (2015) Why AI doomsayers are like sceptical theists and why it matters, *Minds and Machines* **25**, 231-246.

Darrach, B. (1970) Meet Shaky, the first electronic person, *Life Magazine*, November 20, p 58-68.

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/

Davidson, T. (2023) What a compute-centric framework says about takeoff speeds, Open Philantropy, June 27.

Faggella, D. and Leahy, C. (2025) Slamming the brakes on the AGI arms race, *The Trajectory* podcast, January 10, https://danfaggella.com/leahy1/

Garfinkel, B., Brundage, M., Filan, D., Flynn, C., Luketina, J., Page, M., Sandberg, A., Snyder-Beattie, A. and Tegmark, M. (2017) On the impossibility of supersized machines, https://arxiv.org/abs/1703.10987

Good, I.J. (1965) Speculations concerning the first ultraintelligent machine, *Advances in Computers*, vol 6 (eds Alt, F. and Rubinoff, M.), Academic Press, New York.

Greenblatt, R., Shlegeris, B., Sachan, K. and Roger, F. (2023) AI control: Improving safety despite intentional subversion, https://arxiv.org/abs/2312.06942

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

Guan, M. and 14 others (2024) Deliberative alignment: Reasoning enables safer language models, https://arxiv.org/abs/2412.16339

Gurnee, W. and Tegmark, M. (2023) Language models represent space and time, https://arxiv.org/abs/2310.02207v3

Habryka, O. (2024) OpenAI email archives (from Musk v. Altman and OpenAI blog), *LessWrong*, November 16.

Hadfield-Menell, D., Dragan, A., Abbeel, P. and Russell, S. (2017) The off-switch game, *The AAAI-17 Workshop on AI, Ethics and Society*, https://arxiv.org/abs/1611.08219

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

Häggström, O. (2019) Challenges to the Omohundro-Bostrom framework for AI motivations, *Foresight* **21**, 153-166.

Häggström, O. (2021) AI, orthogonality and the Müller-Cannon instrumental vs general intelligence distinction, https://arxiv.org/abs/2109.07911

Häggström, O. (2022) Artificial general intelligence and the common sense argument, in *Philosophy and Theory of Artificial Intelligence* 2021 (ed. V. Müller), Springer, New York, p 155-160.

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

Häggström, O. (2023b) *Tänkande maskiner: Den artificiella intelligensens genombrott*, 2nd ed., Fri Tanke, Stockholm.

Häggström, O. (2024a) On OpenAI's preparedness framework, *YouTube*, January 5, https://www.youtube.com/watch?v=ilwKQSQkQTU

Häggström, O. (2024b) On the troubled relation between AI ethics and AI safety, https://www.math.chalmers.se/~olleh/AIethicsVSAIsafety.pdf

Hanson, R. (2023) Defy your neck-hairs, *Overcoming Bias*, June 20.

Hashim, S. (2024) OpenAI employee says he was fired for raising security concerns to board, *Transformer* Substack, June 4.

Heaven, W. (2023) Geoffrey Hinton tells us why he's now scared of the tech he helped build, *MIT Technology Review*, May 2.

Hendrycks, D. (2023) Natural selection favors AIs over humans, https://arxiv.org/abs/2303.16200

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

Herrman, J. (2023) What does it mean that Elon Musk's new AI chatbot is 'anti-woke'?, *Intelligencer*, November 7.

Hobbhawn, M. and Chan, L. (2023) Should we publish mechanistic interpretability research?, *AI Alignment Forum*, April 21.

Hu, K. and Kai, K. (2024) Exclusive: OpenAI to remove non-profit control and give Sam Altman equity, *Reuters*, September 26.

Isaacson, W. (2023) *Elon Musk*, Simon & Schuster, New York.

Johansson, M. (2024) Setting the AI agenda – Evidence from Sweden in the ChatGPT era, *Chalmers AI Ethics Seminar*, *YouTube*, December 17, https://www.youtube.com/watch?v=saSuVqdHP2g

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

Krakovna, V. (2022) Paradigms of AI alignment: components and enablers, *LessWrong*, June 2.

Kristof, N. (1999) Robokitty, *New York Times Magazine*, August 1, https://www.nytimes.com/1999/08/01/magazine/robokitty.html

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D. and Duvenaud, D. (2025) Gradual disempowerment: Systemic existential risks from incremental AI development, https://arxiv.org/abs/2501.16946

Kurzweil, R. (2005) *The Singularity is Near: When Humans Transcend Biology*, Viking, New York.

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

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, *Nature* **521**, 436-444.

Lee, T., Ball, D. and Cotra, A. (2025) Ajeya Cotra on AI safety and the future of humanity, *AI Summer* podcast, January 16.

Leike, J. (2023) Self-exfiltration is a key dangerous capability, *Musings on the Alignment Problem* Substack, September 13.

Leike, J. (2024) "But over the past years, safety culture…", *Twitter X*, May 17, https://x.com/janleike/status/1791498184671605209

Leike, J. and Sutskever, I. (2023) Introducing Superalignment, OpenAI, https://openai.com/index/introducing-superalignment/

Lermen, S. and Ladish, J. (2023) LoRA fine-tuning efficiently undoes safety training from Llama 2-Chat 70B, *LessWrong*, October 12.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O. and Hendrycks, D. (2025) Utility engineering: Analyzing and controlling emergent value systems in AIs, https://arxiv.org/pdf/2502.08640

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C. (1955) A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf

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

Meta (2025) Frontier AI Framework, February 3, https://ai.meta.com/static-resource/meta-frontier-ai-framework/

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

Miles, R. (2024) AI ruined my year, YouTube, June 1, https://www.youtube.com/watch?v=2ziuPUeewK0

Mowshowitz, Z. (2024a) AI and the Technological Richter Scale, *Don't Worry About the Vase* Substack, September 4.

Mowshowitz, Z. (2024b) AI #90: The wall, *Don't Worry About the Vase* Substack, November 14.

Mowshowitz, Z. (2025a) DeepSeek panic at the app store, *Don't Worry About the Vase* Substack, January 27.

Mowshowitz, Z. (2025b) o3-mini Early Days and the OpenAI AMA, *Don't Worry About the Vase* Substack, February 3.

Mowshowitz, Z. (2025c) We're in Deep Research, *Don't Worry About the Vase* Substack, February 4.

Mowshowitz, Z. (2025d) On the Meta and DeepMind Safety Frameworks, *Don't Worry About the Vase* Substack, February 7.

Müller, V. and Cannon, M. (2021) Existential risk from AI and orthogonality: can we have it both ways?, *Ratio*, https://doi.org/10.1111/rati.12320

Nanda, N., Lee, A. and Wattenberg, M. (2023) Emergent linear representations in world models of self-supervised sequence models, https://arxiv.org/abs/2309.00941

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

OpenAI (2015) *Introducing OpenAI*, December 15, https://openai.com/index/introducing-openai/

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.pdf

Ord, T. (2025) Inference scaling reshapes AI governance, February 12, https://www.tobyord.com/writing/inference-scaling-reshapes-ai-governance

Park, P., Goldstein, S., O'Gara, A., Chen, M. and Hendrycks, D. (2024) AI deception: A survey of examples, risks, and potential solutions, *Patterns* **5**(5).

Phan, L. and 700+ others (2025) Humanity's Last Exam, https://arxiv.org/abs/2501.14249

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

Russell, S. (2025) Closing address at the IASEAI'25 Safe and Ethical AI Conference, Paris, February 7, beginning 03:35:50 into the video at https://oecdtv.webtv-solution.com/3ad1563dc6d988498151b9effd6b6bf1/or/IASEAI-25-Safe-Ethical-AI-Conference-Plenary-.html

Samuel, S., Piper, K. and Matthews, D. (2024) California's governor has vetoed a historic AI safety bill, *Vox*, September 29.

Segerie, C. (2023) Compendium of problems with RLHF, *LessWrong*, January 29.

Shapira, L. and Critch, A. (2024) Will AI extinction be fast or slow?, *Doom Debates*, November 16.

Shear, E. (2025) "I think it's actually reductio ad...", *Twitter X*, February 9, https://x.com/eshear/status/1888669979719114986

Silver, N. (2024) *On the Edge: The Art of Risking Everything,* Penguin Press, New York.

Soares, N. (2022) On how various plans miss the hard bits of the alignment challenge, *LessWrong*, July 12.

Soares, N., Fallenstein, B., Armstrong, S. and Yudkowsky, E. (2015) Corrigibility, *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Sutskever, I. (2024) Sequence to sequence learning with neural networks: what a decade, talk at *NeurIPS 2024*, https://www.youtube.com/watch?v=sJE8qZmQWGU

Tegmark, M. (2024) The hopium wars: the AGI entente delusion, *LessWrong*, October 13.

Templeton, A. and 25 others (2024) Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Turing, A. (1951) Intelligent machinery, a heretical theory, reprinted in *Philosphia Mathematica* **4** (1996), 256-260.

Vance, JD (2025) Transcript of JD Vance remarks at Paris AI Summit, *GitHub*, February 11, https://gist.github.com/lmmx/b373b9819318d014adfdc32182ab17ff

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. and Polosukhin, I. (2017) Attention is all you need, *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA.

Wiblin, R., Harris, K. and Chan Loui, R. (2024) Rose Chan Loui on OpenAI's gambit to ditch its nonprofit, *80,000 Hours Podcast*, November 27.

Wiener, N. (1960) Some moral and technical consequences of automation, *Science* **131**, 1355-1358.

Yampolskiy, R. (2024) *AI: Unexplainable, Unpredictable, Uncontrollable*, Chapman and Hall/CRC, London.

Yudkowsky, E. (2008a) Cognitive biases potentially affecting judgment of global risks, in *Global Catastrophic Risks* (eds N. Bostrom and M. Ćirković), Oxford University Press, Oxford, p 91-119.

Yudkowsky, E. (2008b) 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. (2009) Value is fragile, *LessWrong*, January 29.

Yudkowsky, E. (2013) Intelligent explosion microeconomics, Machine Intelligence Research Institute, https://intelligence.org/files/IEM.pdf

Yudkowsky, E. (2022) AGI ruin: A list of lethalities, *LessWrong*, June 6.

Yudkowsky, E. (2025) The Sun is big, but superintelligences will not spare Earth a little sunlight, Machine Intelligence Research Institute, https://intelligence.org/2025/02/06/the-sun-is-big-but-superintelligences-will-not-spare-earth-a-little-sunlight/