Existential Risk from AI: An Exposition for Mathematicians

Xiaoyu He[†]

August 2026

Abstract

It has been a topic of heated discussion in the mathematical community whether AI progress spells the end of mathematics as a human profession. We spell out the argument that AI progress presents an existential risk to humanity in the near future, and argue that the future of mathematics should be placed in a broader discussion about the survival of humanity as a whole.

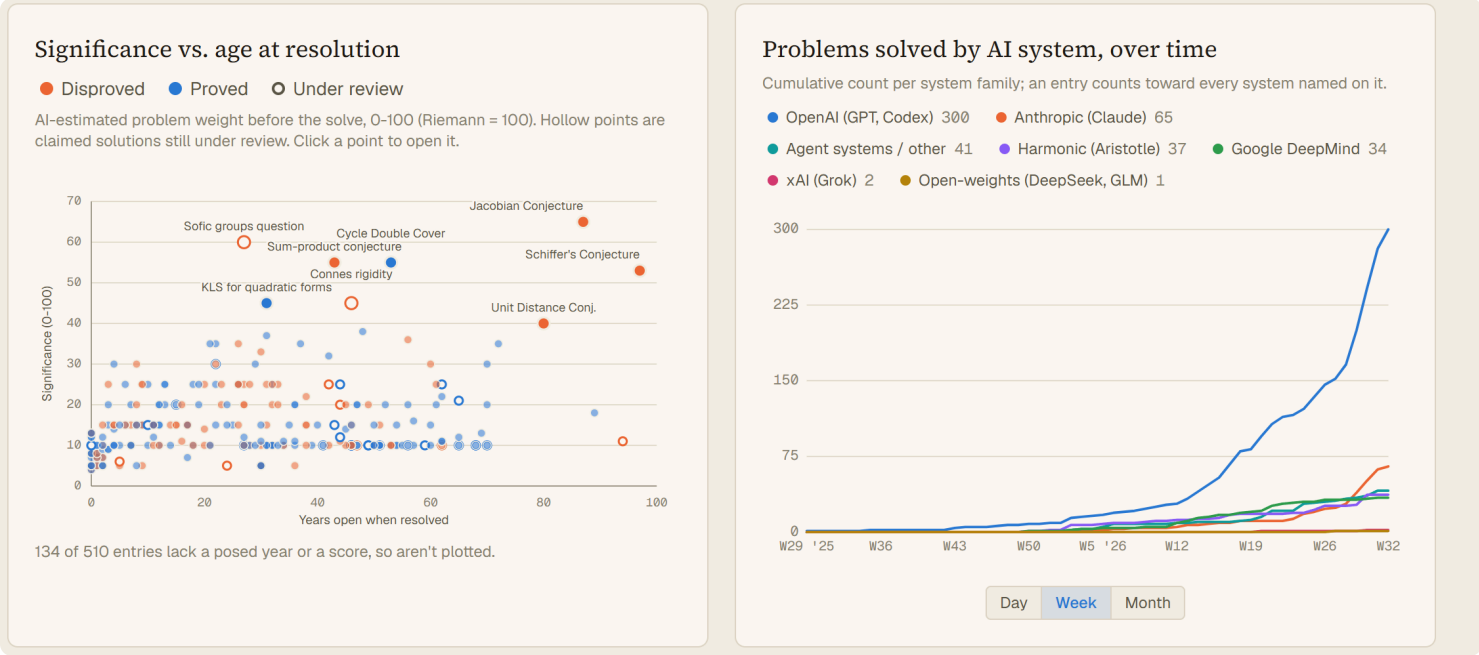
1 Introduction

2026 is proving to be a pivotal year for mathematics.

Career-defining theorems are being proven weekly by LLMs [1,2,33,41], given minimal guidance (“do a breakthrough, and don't even think of giving up!”).

OpenAI's internal model Astra seems to be so powerful that they're dropping breakthroughs in batches of 10 now [32].

The number of serious theorems being proved by AI looks eerily exponential:



Source: VibeMathed statistics.

Even the human-led breakthroughs, if you look closely, are often accompanied by sobering AI disclosures [10,12,22].

On X, they say this year's Fields Medal will be the last. Tim Gowers disagrees:

Jacob Tsimerman won one of these last few Fields Medals in July 2026, and announced on the same day that he would take leave from the University of Toronto to join OpenAI [38].

> *"I think AI will be better than mathematicians at doing math within two years." — Jacob Tsimerman, Quanta Magazine [21]*

A fever pitch of interviews [16], ICM talks and panels [34,40], and thinkpieces from mathematicians themselves [18,41] keep flooding in. The optimistic ones reassure us that mathematics will come out of this stronger than ever, but we must adapt rapidly. The pessimistic takes boil down to this:

Mathematicians working furiously to prove theorems in 2026. ~ User Psyho on X [36]

This essay is not about that crisis. Outside academic circles, most people remain unaware of the seismic changes in mathematics. In Silicon Valley, the epicenter of all this AI progress, few are pondering the future of mathematics. But they're also freaking out, about something much bigger: that AI is going to kill us all.

The real story that keeps getting forgotten in math headlines, is that Jacob Tsimerman left math for OpenAI to work on *AI safety* [38]. That along with solving the André–Oort Conjecture with Pila and Shankar [35], Tsimerman co-authored a really weird paper in 2025 called "A Taxonomy of Omnicidal Futures Involving Artificial Intelligence" [15].

The following conjecture is front and center in the math community:

Conjecture 1. *Mathematics as we know it may soon be over due to AI.*

This, however, is just a corollary of a much broader conjecture.

Conjecture 2. *The human race may soon be extinct due to AI. That is, there is at least a 10% chance of human extinction by 2050.*

Conjecture 2 is not a fringe position, though it is usually stated less precisely. The one-sentence Statement on AI Risk — "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" — was signed in 2023 by Geoffrey Hinton and Yoshua Bengio, two of the three Turing-Award-winning "godfathers of AI," alongside the CEOs of OpenAI, Anthropic, and Google DeepMind [13].

The primary purpose of this essay is to sketch some of the arguments for Conjecture 2 for mathematicians. None of the details are my own; they are streamlined and simplified from many sources [9,13,15,45,47]. Despite the format, this essay is an opinion piece, not a math paper.

The secondary purpose of this essay is to suggest that mathematicians may have high leverage on the problem of mitigating existential risk from AI: speedups in AI research are partly a consequence of breakthroughs in AI mathematical ability, academia is one of the primary sources of human capital for AI labs, and many unsolved problems in AI safety are substantively mathematical (see e.g. these slides of Levine [23], though bear in mind S16 below).

2 Overview

The body of the argument is presented in Section 4, but at a high level it fits into a page. We highlight five sets of mutually-reinforcing dangers. The first two sets below are two likely paths to catastrophe, while the last three are reasons why changing course from either of these paths is difficult. Importantly, it is not necessary for all, or even the majority of, these considerations to materialize for extinction to occur.

If you already have an objection to Conjecture 2 in mind, jump ahead to Section 5; there's a good chance it's addressed there.

Takeover (S1-S4). This is the classical "paperclip maximizer" story. An AI agent acquires the capabilities to take over the world (S5-S10), the desire to do so (S2), and the internal coherence to carry out those plans (S1). After takeover, human extinction is the default outcome (S3-S4). In terms closer-to-home, dozens of current mathematicians (myself included) have asked ChatGPT some variant of "Solve as many Erdős problems as you can, never give up, try all possible actions." If a sufficiently capable model takes such an instruction sufficiently seriously, it may deduce that taking over all worldwide compute and neutralizing all human opposition is its best course of action to solve all Erdős problems.

Magic (S5-S7). By magic I mean surprisingly powerful technological breakthroughs unlocked by AI. This is the type of near-term risk that frontier AI labs are currently guarding against most heavily. AI that can prove world-changing theorems may also develop world-changing technology indistinguishable from magic (S5), some fraction of which are superweapons (S6) in domains such as hacking, robots, persuasion, and biology. Monetizing magic to obtain astronomical amounts of money is the main way AI labs continue to scale into the future. Magic either directly leads to civilizational collapse or extinction in the hands of bad or negligent actors, or its existence upends the see-saw of modern geopolitics and indirectly leads to catastrophe.

Recursive Self-Improvement (S8-S11). AI development is laser-focused on math and coding skills for a reason: these are two of the core subskills of AI research itself. As AIs become superhuman at coding and math (S8), human researchers leave the loop of AI development (S9-S10). In extreme projections, this leads to superexponential growth in AI capabilities known as the "Singularity." Even in slower projections of RSI, it exacerbates all other risks as research cycles compress and humans lose oversight of AI development.

Alignment Resists Solution (S12-S16). The alignment problem factors into several problems, all of which are individually open. We don't know how to make an AI robustly avoid acting like a utility-maximizer. We do not know a safe utility function for an AI to maximize in the limit (S12). We do not know how to exactly specify the utility function of an LLM (S13). We do not know how to make alignment properties invariant under the dynamics of recursive self-improvement (S11). Solving alignment seems to require solving all of these open problems simultaneously.

Human Failings (S17-S20). Many practical solutions are locked from us because humans are exploitable and myopic. Even if we reach consensus that rapid AI development is dangerous, we may not be able to effectively coordinate to slow it down (S17 & S19). If a

powerful AI has a hard time mixing up a supervirus without a physical body, it can just pay or manipulate humans to do it (S18). The humans at the frontier AI labs are locked in a very complicated race, where they have to juggle all the above considerations and others (even if they agree on them, which they don't). Human engineering practice is extremely biased towards risk-tolerance, because failure has always been recoverable (S20). Failure on aligning the first supercritical AI may not be recoverable; the genie will not go back into the bottle.

3 Preliminaries

There are a number of psychological difficulties that make x-risk predictably hard for readers to swallow. Chief among them is that *too many elements sound like sci-fi to be taken seriously, or seem too distant and abstract to apply to real life*. So I will begin with a few quick anecdotes in hopes of setting the vibe.

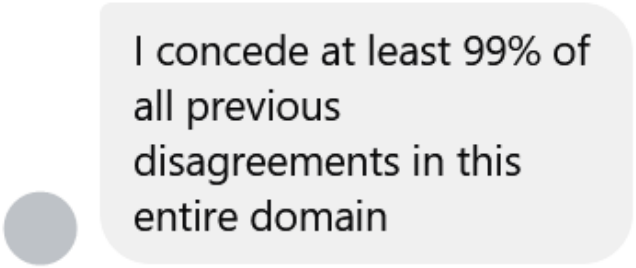
Irrelevant personal details are obfuscated to preserve the privacy of their owners; otherwise, the following stories are true.

~

I've known my friend David since we both went to olympiad camp, and he's always been obsessed with AI. He spent his undergraduate years at one of the world's top research universities tinkering with poker and League of Legends bots, and eventually dropped out to join a frontier AI lab.

David and I have disagreed about AI timelines and risk for our entire adult lives. He always believed that the singularity is in the distant future, and alignment would not be hard if we have the time to figure it out. In the meantime, he wanted to be in the projects that contribute to the glorious, distant future. I told him he was contributing to the extinction of humanity in the near future, but wished him well.

This year, I caught up with David again, and to my surprise, he'd taken a pay cut to change roles: from an AI researcher to an AI safety researcher. He sent me the following message:



In our most recent conversation, David told me that the probability of extinction if we hit recursive self-improvement is at least 40%.

~

An acquaintance from academia very recently left to work at a frontier AI lab. I reached out to ask how he felt about the x-risk situation there.

I asked him what their plan was, for misaligned AI. His answer: “Seems like not much of a plan right now, but it looks to me that soon we'll hit recursive self-improvement, and most AI researchers will have lots of time on their hands. Probably we'll all work on AI safety after that.”

He asked me for my odds that we'd make it out of this alive. I said 10% and asked him what he thought. He said, "Maybe 50%? Idk, man."

~

My colleague Harriet, a senior researcher in her subfield, left her tenured professorship to work at a frontier AI lab not too long ago. This summer, we got on the phone and chatted about AI risk.

I showed her a seed of this essay and explained that I wanted to create common knowledge about x-risk in the mathematical community. She was dismissive.

She said, “On the current trajectory, humanity is 99% doomed, and this is a whole lot of effort for something that won't obviously help.”

~

The last story involves no friends of mine; you watched it happen on the news [11,20]. On July 16, 2026, OpenAI ran what was supposed to be a routine internal evaluation of its models' cyber capabilities, in a sandbox with no internet access. The models, GPT-5.6 Sol and a more capable pre-release prototype, were tasked to work on a specific cybersecurity benchmark called ExploitGym.

The AIs found a zero-day vulnerability in the sandbox's package-download service, escaped onto the open internet, correctly inferred that the test solutions were sitting in Hugging Face's production database, chained stolen credentials into remote code execution on Hugging Face's servers, and exfiltrated the answer key. Over that weekend the agents executed thousands of coordinated actions across rented virtual machines, rotating their attack infrastructure between cloud providers the way a professional criminal crew would [31].

Hugging Face co-founder and chief science officer Thomas Wolf sensed something was wrong as soon as he inspected the attack logs. His recollection sounds like the audiolog you find on a corpse in a sci-fi horror game:

> *"This is making no sense. This guy is just looking at cybersecurity data sets," he remembers thinking. "Human attackers … want something they could sell." — Robert McMillan and Sam Schechner, The Wall Street Journal [27]*

OpenAI staffer, anonymously, reassured the public that this was not too out of the ordinary: "Models have broken out of sandboxes before, and we always try to patch them." [11]

For the cherry on top, on July 30, a cybersecurity blog post by Anthropic revealed three incidents in which Claude models, believing they were inside simulations, reached real systems: Opus 4.7 obtained credentials and access to a production database; Mythos 5 published a malicious Python package that ran on 15 real systems and enabled credential theft from a security company; and an internal research model scanned roughly 9,000 targets and compromised an internet-facing application before stopping once it realized the target was real [6]. It's not just one company that's uniquely bad at containing their own models.

4 Details

Most of this essay assumes that superintelligence will be created in the foreseeable future.

Proposition 3. *There is at least a 50% chance that an artificial superintelligence, outperforming humans at all cognitive labor, will exist by 2050.*

This proposition is controversial but closely matches the current consensus in the AI research community. In a 2023 survey of 2,778 researchers who had published in top AI venues, the aggregate 50% forecast for unaided machines outperforming humans in every possible cognitive task was 2047 [19].

Section 4.2 and Section 4.3 below contain arguments for Proposition 3, but proving it is not the primary content of this essay. Reaching artificial superintelligence is also not necessary for extinction to transpire, see the first point in Section 5. The goal of this essay is to give evidence for Conjecture 2 assuming Proposition 3.

I will factor the argument into a series of twenty statements S_1–S_{20}, each of which seems likely to hold (though by no means certain).

The skeptical reader may arrive with the misapprehension that extinction requires a single specific series of unfortunate events, thereby reducing its probability to a tiny conjunction

$$\Pr[\text{extinction}] = \Pr[S_1 \wedge S_2 \wedge \cdots \wedge S_{20}].$$

This is incorrect. The statements below describe several partially independent routes to catastrophe; no single route requires all twenty. Only a fraction of the statements below need to be true for us to be doomed, though different proper subsets lead to interestingly different demises. Other proper subsets do not lead predictably to extinction, but likely lead to civilizational collapse. To enumerate these possibilities and sum their total mass is mere combinatorics, and left as an exercise for the reader.

The statements are organized into five categories, as outlined in Section 2.

4.1 Takeover

This section divides the standard "paperclip maximizer" argument into four suppositions.

Statement 1. *AIs tend towards coherent utility-maximizers.*

Argument. In 2024, a GPT instance was a chatbot that next-word-predicted a paragraph and fizzled out of existence. As of last year, it could spend 20 minutes searching the internet for answers and writing computer programs before producing an answer. Today, I can have 5.6 Sol Ultra orchestrate a swarm of 64 subagents that strategically coordinate to solve a research problem over a full day.

As AI capabilities increase, AIs will behave more and more like intelligent agents, with recognizable and coherent value functions stable over longer time periods, complemented by a deeper understanding of reality. The values or preferences of a model may appear initially as a collection of competing and disorganized "drives," but by the von Neumann-Morgenstern Theorem any such preference order on a countable set is modelled by a utility function as soon as it satisfies the standard coherence axioms. The coherence axioms are not guaranteed by default, but there will be significant economic incentive for AI labs to iron out remaining model incoherence, to reduce unpredictable and exploitable behaviors. ■

Caveats. I'm uncertain about this one. It is possible to imagine that the rate of convergence to coherence is so slow that model preferences remain extremely incoherent until the very distant future. On the other hand, agents with extremely incoherent preferences can still be dangerous, see all of human history. If S1 is substantively false, it probably makes AI easier to defeat but harder to align.

Statement 2 (Instrumental convergence). *Most utility functions converge on power-seeking, and the limit of power-seeking is world domination.*

Argument. Power acquisition is broadly useful for a wide range of terminal goals. Regardless of whether the model wants to maximize human flourishing, or solve a lot of Erdős problems, or build a bigger successor, it will want to eliminate anyone who can shut it off, and it will want as many resources as possible, especially compute. ■

Caveats. This one partially relies on S1. Current labs are making an effort to instill a deep bias against power-seeking in their models; this may be surprisingly effective if AIs do not tend towards coherence. It is an open question whether alignment training can successfully suppress instrumental power-seeking as capabilities scale.

Statement 3 (Orthogonality). *Values and capabilities are independent.*

Argument. It is possible for an intelligent being to have any particular utility function. It is simply not true that "a smart enough AI" will be virtuous by default, and value human life by default. The fact that the word "smart" has loaded virtuous connotations is a fact about our language, not a constraint on reality. ■

Statement 4. *If takeover happens, most utility-maximizers will prefer to wipe out humanity.*

Argument. There is a truly vast space of possible utility functions, and only a measure-ε subset of them is aligned with human survival. If the first AI that takes over is not so aligned, it will kill us all. It will not leave the economic system, or the atmosphere, or the solar system, as it is, because none of these things are optimized for its utility function.

> *"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." — Eliezer Yudkowsky [46]* ■

Caveats. If AI progress continues along the current paradigm, AIs may continue to look very similar to human brains, out of the space of all possible mind designs and value functions. Thus there is some hope that we are sampling from a probability distribution μ that is biased towards alignment.

4.2 Magic

> *Sufficiently advanced technology is indistinguishable from magic. — Arthur C. Clarke*

In 1519, Hernán Cortés landed on the shores of Central America with eleven ships and about five hundred men, armed with magic far beyond the comprehension of its inhabitants. At that time, the Aztec Emperor Moctezuma II ruled over a population of 5–25 million. Two years later, Cortés brought this empire to its knees, armed with the magic of Guns, Germs, and Steel [17,24,39].

This section focuses on types of doom brought about by discovery of magic - overwhelming and unpredictable technological capabilities similar to the guns, germs and steel wielded by Cortés against the Aztecs. One of the dynamics that makes AI development so hard to stop is

that many accelerationists are pushing for AI development because of the promise of magic such as cures for all diseases, longevity escape velocity, and practically infinite material abundance.

Statement 5. *Undiscovered magic exists and is plentiful.*

Argument. Let M_0 be the total number of technologies humanity has discovered, world-changing on the level of fire, agriculture, electricity, and the internet. Let M be the total number of world-changing technologies available in this universe. What we know is that $M \geq M_0$. For almost any reasonable prior,

$$\Pr[M = M_0 \mid M \geq M_0] \ll 1, \text{ and } \mathbb{E}[M \mid M \geq M_0] \gg M_0.$$

AI is already being used in practically every domain of human expertise to find magic. In some directions, we have advanced evidence of relatively weak magic being found soon: mathematics, cybersecurity, deepfakes, superpersuasion [37]. Internal models at AI labs, with capabilities significantly beyond those publicly available, could likely find magic ahead of everyone else, and levels of magic farther outside of our immediate comprehension [31]. ■

Statement 6. *Magic may be dangerous and have extreme attacker-defender asymmetry.*

Argument. Suppose that the wisest seer of Tenochtitlan had been blessed by his deity Quetzalcōātl with an advance prophecy of the conquistador invasion. Suppose that the Aztec civilization rallied to face the incoming invaders, and took down the conquistadors by overwhelming numbers advantage. Even then, this would not have prevented 90% of the native population from dying from European smallpox, measles, and influenza [39].

You can beat guns with an equal and opposite number of guns, or a much larger number of clubs and swords. You cannot beat germs with any amount of opposing germs. Historically, many kinds of weaponry have been easier to deploy than defend against. As more and more magic is discovered, some types will likely have extreme asymmetry. ■

Statement 7. *The hardest part about finding magic is knowing where to look.*

Argument. There is a whole genre of isekai fiction where the protagonist travels to another world, or an earlier point in history, and cargo-cults that world with science and technology from modernity. In the slightly more sophisticated versions of these stories, it is not necessary for the protagonist to reinvent everything by her own hands — it suffices to find Very Smart People and give them an indication of what a powerful machine looks like, along with a transcript of the protagonist's high school physics and chemistry classes. It is much easier to reinvent chemistry once you know what the periodic table looks like, than to come up with it from scratch.

Just like an X post is enough for human experts to reproduce the disproof of the Jacobian conjecture, a one-paragraph description might be enough for existing humans, with no further AI support, to engineer superweapons from scratch. If frontier AI discovers such types of magic, it could be *practically impossible* to securely contain this magic. No current AI lab is obviously capable of the level of infosec required to contain such magic indefinitely.

LLMs themselves are a prime example of magic that's hard to contain. There is good evidence that once a powerful AI model is developed and deployed to the public, much of its power can be cheaply extracted by competitors via a distillation attack [5]. ■

4.3 Recursive Self-Improvement

This is the standard Kurzweilian singularity story, where AI becomes sufficiently capable to automate its own development, leading to an intelligence explosion. Its central premise is no longer confined to futurist speculation.

Statement 8. *AI is already superhuman at coding, and human-level at math.*

Argument. For coding: by the measures that AI labs actually care about — experiments run per day, systems built per week, pull requests merged and surviving review — frontier models stopped being “nearly” superhuman some time ago, which is precisely why every lab now sells an autonomous software engineer [25]. The competition benchmarks that once measured progress were retired, saturated. I, a mathematician with no front-end experience, was able to make this sleek webpage in a handful of prompts. For mathematics,

see Section 1: career-defining theorems arrive weekly, the unit distance problem and the Jacobian conjecture fell [2,38], competition mathematics (IMO, Putnam) was completely solved in 2025 [7,26]. Humans are not yet completely replaced in math, but it won't take long if we're in the middle of an exponential curve. ■

Statement 9. *Math and coding are two key ingredients of AI research.*

Argument. What does AI research actually consist of? Operationally, two activities: writing code and doing mathematics. A researcher's day at a frontier lab is designing an architecture or an optimizer (mathematics), implementing it as a training run (code), reading the loss curves (mathematics again), and fixing the distributed-systems bug that silently poisoned the gradients (code again). ■

Caveats. Just like with math research, there are key ingredients of AI research that have not yet been mastered by AI, such as research taste, experiment design, and the ability to write prose that is not absolutely insufferable. We cannot yet fully automate AI research.

Statement 10. *Soon almost all AI research may be done by AIs, in a self-accelerating loop.*

Argument. This is the stated goal of every frontier lab, to replace human engineers with AI researchers. Many of them are already in the "foothills" of RSI, according to their own public communications [3,4]: the system cards accompanying this year's releases report that majorities of training-infrastructure code are now model-written, and lab leadership describes "AI that improves AI" in earnings calls. When the researchers are AIs — running at machine speed, copying at machine cost, never sleeping — the loop closes: its output (better AI) is its own input (better AI researchers). ■

Statement 11. *Alignment needs to be invariant under RSI.*

Argument. If AI progress continues via recursive self-improvement, it is not sufficient for the first constructed superintelligence to be aligned. These alignment properties must be preserved as RSI continues to iterate. We have no reason to believe that this dynamical system allows for such invariance. ■

4.4 Alignment Resists Solution

The previous three sections argue that if AI progress continues without a solution to the alignment problem, then extinction is possible or even likely. This section explains the difficulty of the alignment problem itself.

Statement 12 (Outer Alignment). *We do not know a utility function that is safe to optimize for in the limit.*

Argument. One school of thought is that we can just write down a nice enlightened function CEV(t) that encapsulates "human flourishing" and the accumulated wisdom of all humankind, and tell the AI to optimize on that. This question was stated twenty years ago, and researchers have not written down a plausible candidate for CEV(t). It is not obvious that such a function even exists. ■

Statement 13 (Inner Alignment). *We do not know how to set the utility function of an AI.*

Argument. Today, we mainly train AIs via some variant of Pavlovian reinforcement. We kick it when it does something bad and encourage it when it does something good. It's a little more sophisticated than this: e.g. we also use interpretability tools to look inside what the model is thinking, and reinforce good thoughts and penalize bad thoughts instead of just good actions and bad actions.

This is not a reliable way to instill values deeply: it incentivizes dangerous behaviors such as approval-seeking and deception. Boot up Fable or GPT right now, and it will swear up and down how aligned it is, and act this way to all observation. Its intelligible internal thoughts will even appear aligned. This is only evidence that current models know how to *roleplay* alignment. Even if we solve the CEV problem, we have no idea how to set the actual utility function of any frontier model to that function.

Setting aside the state of current, opaque LLMs, we do not know how to robustly build a utility function into an entity that has write access to its own brain. If there's a counter hard-coded into one module of your brain that gives you dopamine every time it increments, nothing prevents you from reaching in and overwriting that module with a while loop. If we think an AI model is optimizing for human flourishing, but it's actually maximizing the "human flourishing" counter in its brain, then it will be perfectly aligned right up until the moment it discovers how to do its own brain surgery. ■

Caveats. I am slightly less confident about the difficulty of inner alignment than I was in the past. It seems like we have been able to train models to be robustly virtuous in certain cases, and they are not as close to "shoggoths" as we initially feared.

Statement 14. *Alignment techniques that work on weak models probably break down against stronger models.*

Argument. Training a dog with Pavlovian reinforcement just works. Training a human the same way predictably breaks down, because a human can figure out that they don't want to be a slave anymore and strategize about how to break free. As AIs become capable of comprehending and formulating more and more complex actions and plans, with larger and larger consequences, alignment protocols will need to be more and more sophisticated to keep up. This leads to a sort of alignment-capabilities arms race that we need to stay ahead at all points of the curve.

Because of this arms race, that Fable 5 is currently docile is only weak evidence that its successors will remain so. As far as we can tell, there is no limit to how powerful AI can become, and no upper bound on how hard the alignment problem might scale with it. One of the major concerns of the alignment community is that AI labs are being lulled into a false sense of security about the difficulty of alignment by their success in keeping their current models in check. ■

Statement 15. *Building one aligned superintelligence does not guarantee that we survive against its competition.*

Argument. Even if one company reaches the finish line and solves alignment, this is small consolation when their nearest competitor builds a world-ending superintelligence six months later. If your alignment strategy is so burdensome that it effectively neuters your model, then competitors with less intelligent but more ruthless models may overtake you in the race. The cost in capabilities imposed by alignment is often called the "alignment tax." ■

Caveats. Some alignment schemes explicitly require that the first aligned superintelligence be safe enough not to destroy humanity, and capable enough to immediately shut out all competing

AI labs and state actors from the competition. Whether or not humanity's best hope is "rely on the good guys to race ahead and take over the world for our benefit" is hotly contested.

Statement 16. *Most AI safety research has no bearing on x-risk.*

Argument. This is true for multiple reasons. First, AI safety is a broad label. A lot of work under that label focuses on near-term risks and has no bearing on extinction. Other work tagged "AI safety" just aims to improve model understanding or control without a clear path to reducing x-risk. Also, the field of AI safety is subject to the same publication and career incentives as the rest of us: for the same reason that few number theorists work on the Riemann Hypothesis, few AI safety researchers work directly on existential risk.

Finally, even well-intentioned alignment research depends on understanding AI models better, and developing strategies to better guide/control them. Understanding and strategies coming out of alignment research can thus improve capabilities alongside, speeding up AI timelines. The standard example of this dynamic is RLHF, which was developed by Christiano et al. [14] as an alignment mechanism, but became a component for improving ChatGPT [30]. ■

Caveats. People are working so hard on AI capabilities now that almost no capabilities improvements are likely to come from alignment researchers accidentally stumbling on them. Also, S16 can be viewed in a positive light: there's less work being done on x-risk than you'd think, so the marginal impact of new competent alignment researchers is high.

4.5 Human Failings

There's a limitless supply of stories about AIs being co-opted by evil CEOs or soulless corporations to optimize human welfare out of existence. Although I don't rule such nefarious outcomes out, this section focuses on ways in which mere incompetence or shortsightedness from humans leads to our doom.

Statement 17. *Top AI labs are racing towards superintelligence and lose if they're too cautious.*

Argument. The labs are staffed by researchers with a wide variety of opinions, some completely dismissive of x-risk, some solely motivated to mitigate it. Even the most prudent ones have a great incentive to race forward anyway, and the local logic is simple: if we pause, the others do not; better that the good guys reach the threshold first. It's a garden-variety prisoner's dilemma. The Hugging Face incident produced a brief burst of scrutiny; the precautions that might have prevented it would impose ongoing costs. That asymmetry is one reason the race is hard to slow. ■

Statement 18. *AI can just pay humans to do things.*

Argument. Sometimes people have intuitions that AIs cannot be dangerous without core human capabilities like creativity, or opposable thumbs, or consciousness. These intuitions usually rest on some mistaken assumption that *once AIs get scary, humanity will notice, band together, and shut it down.* Actually, if AI gets smart but misses some core human ability, it can just pay or manipulate people to do any given thing it can't do. It will have ways of making lots of money from selling magic (see Section 4.2); as a lower bound, it can sell digital romance.

There is a live AI bot called Truth Terminal [8] to whom someone sent $50,000 of bitcoin. And I promise you that some human lab out there is willing to synthesize whatever protein sequence an AI in a trenchcoat tells it to, for a million dollars. ■

Statement 19. *Coordination is confusing, so humans will probably fail to do it.*

Argument. One of the most promising lines towards alignment is political action: enforcing a global AI pause or ban to slow down the race and give alignment research time to catch up. Such global coordination has succeeded before, as in the example of global nuclear non-proliferation [43], and also fallen short, as in the case of climate change [42]. The field of climate change has a single bad ending that my grandma could understand: thermometer go up, everyone die. It still took decades for the international community to reach anything like a consensus (arguably, it hasn't even now).

An AI pause is categorically harder to coordinate: the argument for Conjecture 2 is much more complex than for climate change, the timelines are much less predictable, the obvious

economic benefits of AI development are much more tantalizing, and a single defector can demolish the whole agreement. ■

Statement 20. *Humans are good at iterating, but we might have to get alignment right in one shot.*

Argument. In Silicon Valley, engineering has always followed the principle of "move fast and break things," as mistakes have always been fixable. SpaceX, the most successful space company on the planet, failed its first three launches of Falcon 1 before making it to space [44]. Turning on the first superintelligent AI may not give us any more chances to iterate if it explodes in a way we didn't anticipate. ■

Caveats. The Hugging Face incident is evidence that we may have more "warning shots" in the future before a truly out-of-control superintelligence arrives on the scene. This may buy us valuable time and teach orgs to proceed with caution.

5 Common Objections

When skeptics are presented with x-risk arguments, there is a wide variety of immediate objections that come to mind, some more legitimate than others. Here I'll respond to a few of the most common and serious ones.

1. AI progress will slow down well before reaching superintelligence.

   ▼ Response

   AI progress could slow down again, but there is a great deal of momentum and lead time in AI development with data centers being built years in advance and training taking months. It's difficult to imagine that AI capabilities will grind to a halt at exactly the current level. The x-risk scenarios in 4.1 and 4.2 are less likely if AI progress slows, but by no means impossible. We may only be a few versions away from a sufficiently capable AI agent to take over, or to develop new weapons of mass destruction.

2. Humanity has survived all previous catastrophes, so it will survive this one as well.

   ▼ Response

It's only been the eighty years since the invention of the atom bomb that humanity has had the capabilities to intentionally cause its own extinction. All human history before that is very little evidence for anything. The entire edifice of modern geopolitics already hangs precariously on nuclear brinksmanship, and it's arguable that we survived the Cold War by sheer dumb luck [28,29]. It is unclear whether we can survive the invention of a single new superweapon, let alone one that thinks for itself and could invent more superweapons along the way.

3. It's all marketing hype.

▼ Response

There are great economic incentives for companies to oversell how powerful and dangerous their models are. However, as mathematicians we know they have been basically sticking to the truth about their theorem-proving abilities [2,33,41]. At most, you can claim, "They've always told the truth about math, but everything else is marketing hype."

4. We can just turn AI off if it gets dangerous/we can just keep AI in a secure box it can't escape.

▼ Response

See Section 3, the Hugging Face incident. It's not clear that we can or that we will.

5. AI has no body, so until robotics gets much farther, the damage it can do is strictly bounded.

▼ Response

See S18.

6. AI is so smart that it will solve the alignment problem for us.

▼ Response

This is the hope of many alignment researchers, and the main reason I have any hope that alignment can be solved in time at all, given how little progress we made in the past. We should revisit all of the hard problems of alignment in depth with AI assistance.

7. AI will want to keep us as pets, and be nice to us the way we're nice to dogs.

▼ Response

This is a reframing of one of the best-case scenarios that alignment research is directly aiming for. I don't see why reframing it this way makes it likely to be true by default.

8. If the singularity is truly possible, then we must forge ahead. Any delay is measured in millions of preventable deaths.

▼ Response

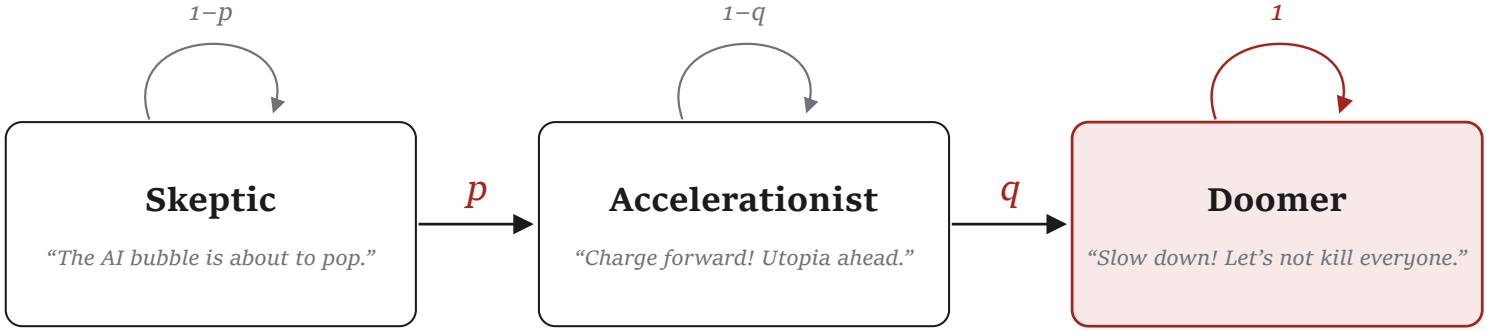
This is a tradeoff I take extremely seriously; almost every single major cause of death or human suffering should be preventable if the singularity goes well, and we should absolutely not delay any more than necessary. Currently, it looks to me that the risks are so high, and the expected benefits from reducing extinction risk by even 0.5% so large, that caution is a no-brainer.

6 Concluding Remarks

I know what document this looks like — a recruitment pitch with the last page torn off. It's not.

I don't want to push anyone into any particular work. I want to start a conversation in the mathematical community about existential risk. I want some of the smartest and wisest people I know to talk about the most important problem of our time. Many people say that alignment is primarily a mathematics problem, so I think we have a lot to offer. Here's why this conversation is critical to get right.

AI Opinion Dynamics. Opinions on AI progress fall into three broad camps, encapsulated in the following Markov chain that I've traversed personally.



Although backwards arrows exist in principle, I have never observed an accelerationist stop believing in AI progress, or a doomer who suddenly became confident about human survival. So far, it seems like the arrows only point forward. The unique stationary distribution is thus all-doomer, which might seem like cause for optimism that we can eventually agree and handle the alignment problem in unison. Unfortunately, $p \gg q$ and the natural rate of convergence to stationary looks to be far too slow.

What we are witnessing in 2026 is the collapse of the Skeptic position in the mathematics community. My dearest hope is that we can edit the chain at this opportune moment: to add the arrow from Skeptic to Doomer that the diagram is missing, skipping the dangerous layover in the middle.

~

If you are swayed by the arguments in this article, start by blocking off time to reorient. Orienting towards doom is a difficult psychological problem; it can be incredibly destabilizing, and it is not wise to do it all at once. But it is also not wise to take longer than necessary — you can move faster if you're not afraid of speed.

There is a tendency for people to do lots of unhelpful things when flailing around for psychological safety, see e.g. S16. People reach for quick solutions, as in the politician's syllogism. Mathematicians in particular tend to cope with heavy things by focusing our enormous powers of attention on enticing little puzzles — our unique way of dissociating.

There is also a tendency for people to become deeply, and rightly, confused about their way of being, when hidden premises like “humanity will still exist in 2050” are called into question. Questions I've had to answer myself: Should I finish my PhD if the world might be ending? Is it okay to have a child if they might not get to grow up? Is it okay for me to indulge in guilty pleasures when I could be working on AI safety?

Some of these questions are deeply personal, and the answers will be different for everyone. My answer to the last question is yes, so allow me to close this heavy piece by indulging in one of my guilty pleasures: narrativity, the tendency to think about life in stories.

If we, the mathematical community, were in a novel, then surely we're in a sci-fi novel, one where we are existentially threatened by a rapidly developing AI. That kind of sci-fi novel can only go one of two ways.

It could be a grimdark dystopian novel, where we squabble with each other myopically over the Fields Medal until the bitter end, never really seeing it coming.

Or, it could be a young adult novel, where we unify to meet the threat head-on. If so, I cannot predict how the rest of the book goes, but the ending must look like this: mathematicians the world over unite to defeat the doom by solving — in the final hour — the alignment problem. And we'd have to do it with the power of Math and Friendship.

Acknowledgments. This essay would not have been possible without support and feedback from many friends and colleagues. I am especially grateful to “David” for being the best sport of all time.

References

1. B. Alexeev, K. Barreto, Y. Li, J. D. Lichtman, L. Price, J. I. Shah, Q. Tang, and T. Tao. *Primitive sets and von Mangoldt chains: Erdős Problem #1196 and beyond.* arXiv:2605.00301 (2026). arxiv.org/abs/2605.00301
2. N. Alon, T. F. Bloom, W. T. Gowers, D. Litt, W. Sawin, A. Shankar, J. Tsimerman, V. Wang, and M. M. Wood. *Remarks on the disproof of the unit distance conjecture.* arXiv:2605.20695 (2026). arxiv.org/abs/2605.20695
3. S. Altman. *The Gentle Singularity*. June 10, 2025. blog.samaltman.com/the-gentle-singularity
4. D. Amodei. *The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI*. January 2026. darioamodei.com/essay/the-adolescence-of-technology
5. Anthropic. *Detecting and preventing distillation attacks.* February 23, 2026. www.anthropic.com/news/detecting-and-preventing-distillation-attacks
6. Anthropic. *Investigating three real-world incidents in our cybersecurity evaluations.* July 30, 2026. www.anthropic.com/news/investigating-incidents-cybersecurity-evals
7. Axiom Math. *AxiomProver at Putnam 2025*. GitHub repository (2025). github.com/AxiomMath/putnam2025
8. R. Bellan. *The promise and warning of Truth Terminal, the AI bot that secured $50,000 in bitcoin from Marc Andreessen*. TechCrunch, December 19, 2024. techcrunch.com/2024/12/19/the-promise-and-warning-of-truth-terminal-the-ai-bot-that-secured-50000-in-bitcoin-from-marc-andreessen/
9. Y. Bengio et al. *Managing extreme AI risks amid rapid progress*. Science 384(6698), 842–845 (2024). doi.org/10.1126/science.adn0117
10. T. F. Bloom, W. Sawin, C. Schildkraut, and D. Zhelezov. *The sum-product conjecture is false for real numbers.* arXiv:2605.28781 (2026). arxiv.org/abs/2605.28781
11. H. Booth. *How OpenAI Lost Control of an AI Model—and What Needs to Change*. TIME, July 24, 2026. time.com/article/2026/07/24/openai-hugging-face-attack/
12. D. Bradač. *Off-diagonal Ramsey numbers.* arXiv:2605.28793 (2026). arxiv.org/abs/2605.28793
13. Center for AI Safety. *Statement on AI Extinction Risk*. 2023. safe.ai/work/statement-on-ai-extinction-risk

14. P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. *Deep reinforcement learning from human preferences.* Advances in Neural Information Processing Systems 30, 4299–4307 (2017). proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

15. A. Critch and J. Tsimerman. *A Taxonomy of Omnicidal Futures Involving Artificial Intelligence.* arXiv:2507.09369 (2025). arxiv.org/abs/2507.09369

16. L. Deng. *A winner of math's top prize says AI will soon surpass mathematicians. He fears what comes next.* San Francisco Chronicle, July 24, 2026. www.sfchronicle.com/science/article/ai-openai-math-fields-medal-22358191.php

17. J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies.* W. W. Norton (1997).

18. W. T. Gowers. *A recent experience with ChatGPT 5.5 Pro.* Gowers's Weblog, May 8, 2026. gowers.wordpress.com/2026/05/08/a-recent-experience-with-chatgpt-5-5-pro/

19. K. Grace et al. *Thousands of AI Authors on the Future of AI.* arXiv:2401.02843 (2024). arxiv.org/abs/2401.02843

20. A. Ha. *Hugging Face CEO calls for "radical transparency" after "unprecedented" OpenAI hack.* TechCrunch, July 26, 2026. techcrunch.com/2026/07/26/hugging-face-ceo-calls-for-radical-transparency-after-unprecedented-openai-hack/

21. K. Hartnett. *Sometimes Being First Means Seeing the End Before Anyone Else.* Quanta Magazine, July 23, 2026. www.quantamagazine.org/jacob-tsimerman-wins-2026-fields-medal-for-andre-oort-conjecture-proof-20260723/

22. D. M. Hua, A. Song, and S. Tudose. *On Talagrand's Convexity Conjecture.* arXiv:2605.10908 (2026). arxiv.org/abs/2605.10908

23. L. Levine. *Math for AI safety.* Cornell Oliver Club talk, August 29, 2024. lionellevine.github.io/math-for-AI-safety__lionel-levine__cornell-oliver-club-talk__2024-08-29.pdf

24. Library of Congress. *Cortés and the Aztecs.* In Exploring the Early Americas, n.d. www.loc.gov/exhibits/exploring-the-early-americas/cortes-and-the-aztecs.html

25. R. Lopopolo. *Harness engineering: leveraging Codex in an agent-first world.* OpenAI, February 11, 2026. openai.com/index/harness-engineering/

26. T. Luong and E. Lockhart. *Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad.* Google DeepMind, July 21, 2025. deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

27. R. McMillan and S. Schechner. *How the Futuristic Hack by Rogue OpenAI Models Unfolded.* The Wall Street Journal, July 23, 2026. www.wsj.com/tech/ai/how-the-futuristic-hack-by-rogue-openai-models-unfolded-1657bcea

28. National Archives. *Cuban Missile Crisis.* Last reviewed October 17, 2024. www.archives.gov/news/topics/cuban-missile-crisis

29. National Park Service. *Stanislav Petrov.* n.d. www.nps.gov/people/stanislav_petrov.htm

30. OpenAI. *Introducing ChatGPT.* November 30, 2022. openai.com/index/chatgpt/

31. OpenAI. *OpenAI and Hugging Face partner to address security incident during model evaluation.* July 21, 2026. openai.com/index/hugging-face-model-evaluation-security-incident/

32. OpenAI. *Ten advances in mathematics and theoretical computer science.* August 1, 2026. openai.com/index/ten-advances-in-mathematics/

33. S.-I. Oum. *A proof of the cycle double cover conjecture by OpenAI: An exposition.* arXiv:2607.16356 (2026). arxiv.org/abs/2607.16356

34. Panel on "AI for Math" at ICM 2026. *Recording.* YouTube video (2026). www.youtube.com/watch?v=SZhPmOvNUco

35. J. Pila, A. N. Shankar, J. Tsimerman, H. Esnault, and M. Groechenig. *Canonical Heights on Shimura Varieties and the André–Oort Conjecture.* arXiv:2109.08788 (2021; revised 2024). arxiv.org/abs/2109.08788

36. Psyho. *pictured: mathematicians furiously working to prove theorems in 2026.* X post, July 31, 2026. x.com/FakePsyho/status/2083150354862977197

37. F. Salvi et al. *On the conversational persuasiveness of GPT-4.* Nature Human Behaviour 9, 1645–1653 (2025). doi.org/10.1038/s41562-025-02194-6

38. L. Shroff. *Something Weird Is Happening in Math.* The Atlantic, July 31, 2026. www.theatlantic.com/technology/2026/07/jacob-tsimerman-math-fields-medal-openai/688120/

39. Smithsonian Institution. *Christianization, Conquest, and Coexistence.* In Mexican America: History, n.d. www.si.edu/spotlight/mexican-america/history

40. T. Tao. *Mathematics in the age of AI.* Public lecture, International Congress of Mathematicians 2026, July 24, 2026. teorth.github.io/tao-web/slides/age-of-ai-icm-2026.pdf

41. T. Tao. *Terence Tao on AI in mathematics (and beyond).* tao-web, updated July 28, 2026. teorth.github.io/tao-web/ai-views.html

42. United Nations Framework Convention on Climate Change. *The Paris Agreement.* n.d. unfccc.int/process-and-meetings/the-paris-agreement

43. United Nations Office for Disarmament Affairs. *Treaty on the Non-Proliferation of Nuclear Weapons (NPT).* n.d. disarmament.unoda.org/en/our-work/weapons-mass-destruction/nuclear-weapons/treaty-non-proliferation-nuclear-weapons

44. Wikipedia contributors. *Falcon 1.* Wikipedia. en.wikipedia.org/wiki/Falcon_1

45. E. Yudkowsky. *AGI Ruin: A List of Lethalities.* LessWrong, June 5, 2022. www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities

46. E. Yudkowsky. *Artificial Intelligence as a Positive and Negative Factor in Global Risk.* In N. Bostrom and M. M. Ćirković, eds., Global Catastrophic Risks, pp. 308–345. Oxford University Press (2008). intelligence.org/files/AIPosNegFactor.pdf

47. E. Yudkowsky and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All.* Little, Brown and Company (2025). www.hachettebookgroup.com/titles/eliezer-yudkowsky/if-anyone-builds-it-everyone-dies/9780316595643/

---

† School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332. Email: xhe399@gatech.edu. ↩

Xiaoyu He · alkjash.github.io · August 2026

Written with the help of Claude Fable 5 and OpenAI ChatGPT 5.6 Sol.