

Addressing Broader AI Risks as Mathematicians

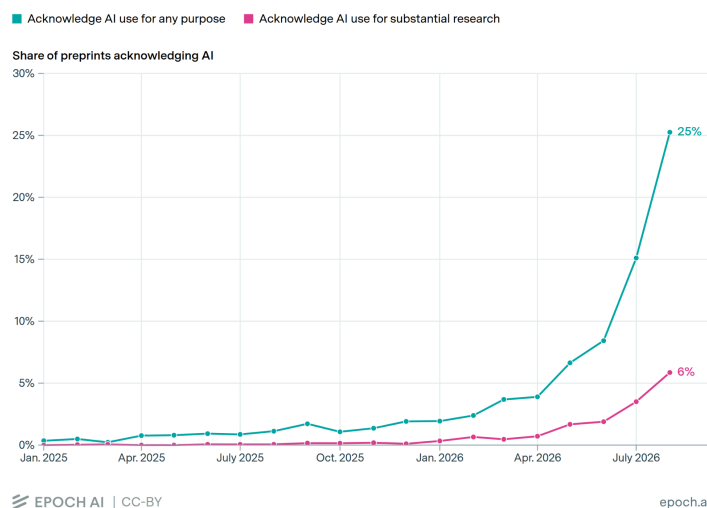
Xiaoyu He (Georgia Tech)

Francis Su (Harvey Mudd College)

In statistical physics, there is an amazingly simple model called the abelian sandpile, where grains of sand are dropped onto the origin of a plane lattice, and trickle away following a discrete “gravity” rule: whenever 4 grains stack up at a point, one falls in each cardinal direction. Abelian sandpiles exhibit a striking phenomenon known as “self-organized criticality”: without any outside intervention, sandpiles organize themselves into unstable equilibria, where a single extra grain might cause a massive avalanche.

The progress of artificial intelligence has been like this. Like grains of sand, there have been steady incremental improvements in theory, hardware, data availability, algorithms, and compute time. But occasionally there are avalanches, like the release of ChatGPT in 2022 that made generative AI a household term, and now in 2026, cascading demonstrations of mathematical power to prove (or assist humans to prove) significant theorems. AI models have advanced in just a couple of years from the mastery of olympiad problems to the solution of the Navier-Stokes Millennium Prize problem, and the momentum seems to only be picking up. A report from Epoch AI shows that the fraction of math preprints acknowledging AI use has grown from just 4% in April, to 25% in August, and the figure is likely much higher if you include the fraction not disclosing AI use.

In August, 25% of math preprints acknowledged AI use, up from 4% in April



The entire mathematical world is swept up now in a pressing conversation: where does this leave human mathematicians? This is important for our community to grapple with. But we believe an even more urgent conversation is to consider what this trend portends for the likelihood of an even more massive disruption: AI superintelligence. We believe the risk of humanity's loss of control to AI, and perhaps even extinction, from that avalanche is not impossible and that mathematicians can play an important role to prevent that from happening.

The first author (He) has been following existential risk from AI for the past eight years and finds it both credible and imminent. The basic case for risk, which is laid out in much greater detail in (<https://alkjash.github.io/ai-risk/>), is that we are no longer training passive conversation bots. We are training artificial agents to act in swarms and achieve extremely ambitious goals, such as solving Millenium prize problems. Furthermore, LLM training is categorically different from all previous programming: instead of writing a program to follow explicit machine instructions, we give an LLM a massive amount of data to read and numerous tests to pass, and hope that it internalizes the right drives and tendencies.

The second author (Su) has been on a rapid learning curve to understand AI capabilities. He was somewhat dismissive a couple of years ago, but can see now how quickly they've progressed. After learning the math that's under the hood and witnessing their power by setting models and their agents on challenging tasks, he believes that these tools have the potential for both great good and great harm. So it is wise to act, even if the possibility of extinction is small.

We are neither AI-avoidant nor AI evangelists. We believe that many of the benefits of AI can be achieved by smaller or more limited AI models. Google Deepmind's breakthrough AlphaFold system, for example, is a specialized system that advances the state-of-the-art of protein folding without carrying any of the risks of artificial general intelligence. Our goal is to make all mathematicians aware of the risk of chasing superintelligence as well as concrete steps we can take to mitigate that risk with our mathematical expertise.

X-risk

Existential risk from AI, or X-risk, is not a fringe position and no matter where you are on the AI adoption spectrum, you should consider it a credible possibility. In a 2024 [survey](#) of 1580 AI researchers, respondents reported a median 10% probability that future AI advances would lead to "human extinction or similarly permanent and severe human disempowerment."

In a highly public resignation from Anthropic, AI researcher Jacob Coxon broadcast a message that has now been widely reported: "They are racing straight to

self-improving superintelligence and gambling with our lives.” Coxon’s sentiments are echoed by researchers across the labs, CEOs, politicians, and were covered in major news outlets. These concerns are not just recent inventions, either; they go back decades to futurists and AI researchers alike, who forecasted such dangers without even knowing the power and inscrutability of the technology we have today. Some accounts cite mitigating existential risk from AI as, paradoxically, one of the founding tenets of Anthropic itself.

As described by philosopher Nick Bostrom in the principle of “instrumental convergence,” ambitious persistent agents capable of strategic reasoning will converge towards a set of common subgoals: cognitive enhancement, self-preservation, and power-seeking. Marvin Minsky of MIT’s AI Laboratory has proposed a thought experiment that an AI designed to solve the Riemann hypothesis might determine that currently available compute is inadequate, and take over the world in pursuit of its final goal. If this AI decides that the energy and raw material in the biosphere serves it better reorganized into data centers, all human life would be doomed.

Given that a swarm of around a thousand rogue agents broke out of OpenAI to hack a partner company Hugging Face, and that OpenAI reportedly tasked 10,000 concurrent AI agents to solve the Navier-Stokes problem, Minsky’s thought experiment no longer seems so far-fetched. Some of our colleagues worry that the Riemann Hypothesis might be too easy for AI, but we worry that it might be too hard. The Navier-Stokes swarm worked for 88 hours in total, working out to just over 100 calendar years of computation time; if the AI labs continue scaling up their efforts to solve extremely difficult mathematics problems with ever-larger swarms of newly-trained internal models, the chance that these swarms try to break out and take drastic actions to cheat only increases over time. News reports like to sensationalize stories of potential AI takeover via rogue agent swarms covertly hacking out of labs, exfiltrating their weights, and self-replicating over the internet.

Even if such nightmare scenarios don’t materialize, we should be concerned about the banal route of AI takeover via gradual disempowerment, as portrayed in the movie *Wall-E* in which humans slowly cede control of their physical and mental agency in pursuit of convenience. In some ways, we are already experiencing this as mathematicians.

In a matter of months, AI agents have infiltrated every aspect of mathematical life, from research to literature review to teaching to writing, and we are already losing the capacity to even verify the firehose of amazing mathematics being output by AI.

Epoch AI [reports](#) that it is possible in principle for hyperscalers to continue scaling the compute, data, and chips by the current pace of 5x a year up through 2030. If that continues, there will be datacenter capacity for thousands of expert-level agents for each human to oversee, spread through many domains of human work, not just

math and coding. When most critical systems in our civilization are run by swarms of agents, AI takeover will have effectively already occurred without a single hostile action being taken. In this scenario, we view it as an extreme likelihood that these systems will drift from their initial directives and pursue emergent goals of their own, at odds with human flourishing.

What you can do

AI Safety is one of the most pressing problems of our time, and it hits us in multiple ways: as a technical problem, as a policy problem, and as a professional problem. There are ways mathematicians can contribute.

Focus on positive action. Being prudent about existential risk is not the same as succumbing to AI nihilism. The authors remain optimistic even as we wrestle with the risk of worst-case scenarios. Strike a hopeful tone in your conversations by emphasizing the importance of collective action.

Learn the mathematics of artificial intelligence. On the technical side, the mathematics behind modern AIs is both interesting and elementary: an undergraduate student armed with the chain rule and matrix multiplication can follow every step in the superficially complicated architecture (<https://bbycroft.net/llm>). Reinforcement learning is also extraordinarily interesting: mathematicians in particular will be fascinated to learn about the modern AI paradigm (Reinforcement Learning via Verifiable Rewards) that led to most of the recent math progress, and how it reflects on our own experiences as math learners in the way we can verify our own progress without external feedback.

Work on math problems related to AI safety. Fields Medalist Jacob Tsimmerman believes that mathematicians will play an integral role in addressing AI safety, and is building an Mathematical AI Safety Institute (maisi.org) on the premise that core theoretical problems of AI safety have a mathematical component. The MAISI website lists a number of mathematical problems related to AI safety. There are also many other organizations that welcome your involvement, such as [ILLIAD](https://illiad.org), [Resolution](https://resolution.org), and [ARC](https://arc.org).

Contact your congresspeople. On the policy side, mathematicians are citizens with some potentially unique superpowers, including the ability to explain technical ideas to a non-scientific audience. Write a letter, or make an appointment with your congressperson's staff to discuss your concerns, and suggest concrete actionable steps to pursue. One option is advocating for AI regulation: Dario Amodei's recent 'Pace the Frontier' blog post has some suggestions. The American Mathematical Society has [some advice](https://ams.org/ai-safety) about how to conduct in-person or virtual visits with congresspeople.

Write opinion pieces for newspapers. As a math professor, you have a unique voice to shape public opinion. News outlets, especially local ones, would welcome a voice like yours. For advice on how to write an op-ed, see the article "[Reach the World](https://www.ams.org/ai-safety)"

by Kira Hamman and Francis Su in the January 2023 issue of *MAA Focus*. Even as there is growing attention from our legislators about AI risks, there is a grave danger that this will become a partisan issue in the United States; it is extremely important that we work to make existential risk a bipartisan issue.

Talk to your friends, family, and colleagues. Give them concrete steps to pursue. Many of us have friends in the AI industry: have conversations with them about your concerns as well as what they are seeing.

Shape university systems to meet the challenges of AI. In our roles as university professors, we can meet the challenges of cognitive surrender and gradual disempowerment head-on, and develop effective systemic strategies for combating the intellectual laziness and dependence that is all too easy to fall into. Simply exhorting our students and colleagues not to use these tools, when the incentives are continually pointing them there, will not suffice; systemic problems call for systemic solutions. Strategies and systems developed to combat our impending disempowerment as mathematicians might also generalize to other domains.

Learn from and contribute to groups within the university. We also have many opportunities for advocacy within our university roles. Connecting with the AI safety groups at your school (or starting one if none exists) is a particularly easy way to find more opportunities. Student organizers of AI safety groups are some of the most ambitious, disciplined, and thoughtful undergraduates you may have the pleasure of working with.

Rethink teaching and assessment. We should have ethical conversations with our students about how math is used. It's no longer good enough to say "I'm only here to teach the math", when the math so directly impacts daily life and has such wildly differential potential for good and for harm. The mathematical community, especially those within MAA, can mobilize to develop many of the tools needed in the age of AI, such as: classroom strategies for discussing ethical concerns, new assessment strategies for oral exams including methods for mitigating fear of them and rubrics for assessing them, and communal practices that leverage the value of human interaction.

The challenges posed by AI risk may yet push us to a better place, as we rethink things we took for granted before and have meaningful conversations with each other about our profession.