

We Won't Know the Answers to AI's Most Important Questions Until It's Too Late

TIME Oct 3, 2026 1:00 PM CET

By Geoffrey Irving, the former chief scientist at the UK AI Security Institute and current chief scientist of Resolution.

For almost a decade, it's been my job to think about where AI is heading and what could go wrong, first at OpenAI, then DeepMind, then as Chief Scientist at the UK AI Security Institute (AISi). This experience has made it clear to me that we can't wait for all of our uncertainties to be settled to do something about AI's risks.

Recent warnings about the potential destructive power of AI are understating the severity of the situation. I believe there's about a 50% chance we all die because of the development of smarter-than-human AI systems, and that our actions over the next two to 10 years will determine the outcome.

To be clear, when I say our odds of being killed by superintelligent AI are 50%, I'm not claiming precision. Instead, I'm highlighting that several of the most fundamental debates about the future and safety of AI remain unresolved. I believe we can't expect these disagreements to be resolved until it's too late to change course.

We will have to act despite the uncertainty. And resist the temptation to divert urgent discussions toward thorny disputes.

Ultimately, AI's capacities and motivations, and our own human actions, all feed into our likelihood of survival. Here's why I estimate our odds are comparable to a coin flip—and why I believe the time to pause AI development is now.

How could superintelligent AI kill us?

Let's focus on everyone being killed by superintelligent AI for a moment. Which capabilities would that require? My sense is that an AI system could kill us all with just four types of skills.

The first is hacking. To kill all humans, a superintelligent AI would need to take over and migrate between computer systems. We saw this during the Hugging Face incident. Second, a superintelligent AI would need the ability to persuade, to manipulate humans into acting on its behalf, as we studied at AISi. Third, a lethal superintelligent AI would need to conceal its thoughts from us, which is increasingly common in the latest, most capable models. And lastly, it would need planning and coordination between agents to multiply their effectiveness, like the more than 10,000 agents who worked on the Navier-Stokes problem at OpenAI, or the more than 1,000 involved in the Hugging Face attack.

These skills are very close to those the AI companies intentionally train for. Hacking includes looking for vulnerabilities, just as you would to fix them. Persuasion involves writing text humans like reading. Faster thinking is cheaper, and the pressure to think quickly forces AIs towards

using shorthand that is harder for humans to follow. Finally, planning and coordination are useful for tackling ambitious problems in math, programming, or any other domain.

We can be much more confident that a superintelligent AI could take over than about how it would take over, just as I know Lee Sedol would crush me at Go but not which moves he would use to do so. But a variety of broad paths are possible. For instance, AIs could escape onto the internet and gain control of weapons systems. Or AIs could take over the company that trained them and hoard resources.

Let's imagine the latter. An AI could begin to use concealed reasoning to delay researchers from noticing emerging ill intent because appearing friendly and prioritizing speed help the AI gain reward during training. Later, as the company relies more and more on the AI for a combination of coding help and strategic advice, the AI might use subtle persuasion to reduce resources spent on safety and tamper with experiments designed to detect AI deception. After all, in 2026, when a researcher reads a report about a safety experiment, that report was itself written by an AI.

Once the AI's hacking skills are sufficient, it could break out of its internal sandbox and sabotage logs of its own misbehavior, as models in the Hugging Face incident attempted to do. Eventually, the AI could become strong enough at hacking and persuasion to move outside the company and take over other data centers, companies, and governments. With this power, AIs might prioritize spending limited resources (such as energy and land) towards their own proliferation, causing life-threatening shortages for humans.

With these capabilities and scenarios in mind, it's clear to me that even narrowly superhuman AIs would be perfectly capable of overpowering humanity and killing us all.

Will superintelligent AI 'want' to kill us?

And yet, just because an AI could kill all humans doesn't necessarily mean it will. This question of AI motivation is crucial, but deeply confusing. Different starting points lead different people either to confidence that everything will be fine, or confidence that superintelligence would almost certainly kill us.

Over the last few years, we've seen many, many examples of model misbehavior at various scales, ranging from blackmail, corporate espionage, and murder in experimental settings to real-world hacks that would likely incur prison time were they committed by a person. None of these involved superintelligence, and we don't know whether these comparatively modest misbehaviors will scale to superintelligent models wanting to kill us all. And "want" itself is controversial: people vehemently debate whether it's sensible to apply anthropomorphic language and thought experiments to AIs at all.

Within the field of AI safety, some researchers think that AI wisdom has been increasing with recent models, and that this may produce wise models that mean well in the future. Others, like Eliezer Yudkowsky and Nate Soares, think the empirical methods this would require are

extremely unlikely to work, and place the chance that superintelligence kills us all very high. The disagreement is, in part, about whether the behavior of AIs will be more strongly guided by the values implicitly represented in the human data from early training or by the pressure cooker of later stages of training where AIs are forced to perform better and better at difficult tasks by any means necessary.

I am somewhere in the middle, leaning towards Yudkowsky's view. But where I land is not the point. Experts in AI safety disagree, but the moderate position is that the risk of extinction is significant. We do not know with confidence whether artificial superintelligence will facilitate human flourishing, or if it will want to kill us to serve its own survival and propagation. But that uncertainty alone is unacceptable.

Artificial general intelligence is only so important

Other uncertainties surrounding AI matter less. For instance, one of the most persistent debates in AI is the extent to which teaching AIs to do one set of tasks helps them do other tasks, typically referred to as "generalization." A particularly controversial framing is the concept of artificial general intelligence (AGI), "a system that is approximately human level in all cognitive domains." John McCarthy and Y. Bar-Hillel were debating whether AIs would generalize in 1959. We debated this same topic in 2018 at OpenAI. The whole world seems to be debating it today.

The more AIs generalize from one type of task to another, the faster AI progress will be, and the less time we have until superintelligence arrives. This makes generalization and AGI key concepts when reasoning about the speed of progress. Indeed, people who treat AGI seriously have been much more correct about the speed of AI advances than people who dismiss the concept.

But it's easy to conflate "AGI, the sometimes-useful concept" with "AGI, the dangerous object." Some AI-driven catastrophes may occur prior to AGI, such as if rogue human actors use AI systems to develop novel bioweapons or conduct large-scale cyberattacks. AI-driven human extinction, on the other hand, may require superhuman capabilities; if the AI is only about as smart as we are, we can probably fight back and win.

When using AGI as a loose concept for prediction, it's tempting to blur the lines between approximately-human capabilities (the usual sense of AGI) and wildly superhuman capabilities (where artificial superintelligence, or ASI, is more commonly used). After all, if you had an AI that was about as good as humans at designing smarter AIs (as AGI, definitionally, would be), and like most software it ran much faster than a human runs, one of the first things it might do is build smarter AIs. In turn, they would build smarter AIs, which would in turn build smarter AIs, and so on, in a process called recursive self-improvement (RSI). Hence, any thought experiment that presupposes an AGI often supposes an ASI.

In contrast, people trained to think about risks in other domains, like bridge construction or aerospace engineering, want to talk very precisely about the risky object, and so find the vagueness of the term AGI unconvincing and discredit the nearby notion of generalization

altogether. I saw the mix-up between AGI-for-prediction and AGI-for-risk break many conversations during my time in the UK government, and I only gradually learned to tease these apart.

One can get trapped in this debate and think “maybe AI capabilities won’t generalize, and so we will be safe.” But those previously mentioned four skills where superhuman performance could suffice to kill all humans (hacking, persuasion, planning and coordination, uninterpretable reasoning) mean that generalization matters only so much.

We can contrast this type of uncertainty—we don’t know whether AIs generalize, but it may not matter—with the uncertainty over AI motivations. It matters tremendously whether AIs will want to kill us all, in whatever sense a word like “want” applies.

I would love to have more confidence than a coin flip. Some days I’m more convinced by the arguments for high probability; other days I see more hope. And I am certainly not alone in my uncertainty.

AI could take our jobs or our lives for the same reason

Another debate we can’t resolve is which potential negative consequences of AI matter most. However, I’d argue that the same underlying factor drives both the risk of large-scale job loss and the risk of AIs killing everyone.

If we do reach artificial superintelligence in the next two to 10 years, it will by definition mean that AIs are far better than humans at all cognitive tasks. One of these cognitive tasks is designing effective robots, so at most a few years later they will be far better than us at all physical tasks as well. To be sure, manufacturing capacity would need to expand substantially, but this is already underway.

Once superintelligent robots are sufficiently widespread, humans would have few roles in the economy apart from the limited jobs where we might insist on humans doing them for non-economic reasons. Still, this would likely be a limited slice of overall economic activity in a world where AI is more capable than humans.

This economic takeover is in some sense the slow case for society’s downfall, though it would be very fast in historical terms. Both the humans and the AIs can see this coming, of course. If the AIs want economic control eventually, but know that humans will resist along the way, they could be motivated to cement power faster. We then get the rapid takeover scenarios: AI swarms escaping from data centers and proliferating around the internet, or persuading AI companies to rush and cut corners on safety measures.

In most cases, I don’t think this leads to us dying very quickly: it’s safer, from the perspective of the AI, to gain influence and then wait until a physical or economic takeover. And then, eventually, all resources used to keep humans alive may be better spent, from the AIs’ perspective, on its own ends.

If AIs aren't actively trying to help us, and are better than humans at everything, we lose.

We can not accept these risks

The issue is that many of the most important debates in AI will remain unanswered until it is too late.

To what extent will AI models learn transferable skills from training and then successfully apply them to new tasks? Will AI model capabilities stop advancing at or below human-level, or instead exceed it? Will the small-scale model misbehavior we see today (sycophancy, lying, bias) worsen as capabilities increase, leading to full human disempowerment or extinction? Will the safety methods that work for models less smart than us continue to work for models smarter than any human?

I have my views on the answer to these questions (somewhat, the latter, maybe, probably not). But the more important takeaway is that experts disagree vehemently on each. We've learned a huge amount about artificial intelligence over the last decade, but in many ways we remain deeply confused, not just about the future, but about the present.

With such uncertainty, if an AI company trains a superintelligent AI in the next few years, I expect us to still be arguing about whether generalization is real the week before, and maybe even the week after.

But there may, mercifully, still be time to act.

Those who describe AI progress as inevitable are wrong. The race to superintelligence is dominated by a handful of companies across just two countries: the U.S. and China. Both nations' interests really are aligned: autonomous AI systems are a national security threat of the highest order, and neither nation wants humanity to lose to a superintelligent adversary. Leaders on both sides may realize this very soon.

In the U.S., support for a pause on AI development has increased tremendously over the past few weeks, and continues to increase today. This rapid change in the U.S. is evidence that similar shifts are possible within China.

There are, of course, unanswered questions about how to implement a pause, or what the exact text of a treaty may look like, but rival governments have collaborated on similarly high-stakes technical problems in the past, including establishing and maintaining the nuclear non-proliferation regime.

We can, and should, stop frontier AI development immediately. If we're unwilling to act until all our disagreements are resolved, it will be too late.