The case for taking AI seriously as a threat to humanity

Kelsey Piper

Stephen Hawking has said, “The development of full artificial intelligence could spell the end of the human race.” Elon Musk claims that AI is humanity’s “biggest existential threat.”

That might have people asking: Wait, what? But these grand worries are rooted in research. Along with Hawking and Musk, prominent figures at Oxford and UC Berkeley and many of the researchers working in AI today believe that advanced AI systems, if deployed carelessly, could end all life on earth.

This concern has been raised since the dawn of computing. But it has come into particular focus in recent years, as advances in machine-learning techniques have given us a more concrete understanding of what we can do with AI, what AI can do for (and to) us, and how much we still don’t know.

There are also skeptics. Some of them think advanced AI is so distant that there’s no point in thinking about it now. Others are worried that excessive hype about the power of their field might kill it prematurely. And even among the people who broadly agree that AI poses unique dangers, there are varying takes on what steps make the most sense today.

The conversation about AI is full of confusion, misinformation, and people talking past each other — in large part because we use the word “AI” to refer to so many things. So here’s the big picture on how artificial intelligence might pose a catastrophic threat, in nine questions:

1) What is AI?

Artificial intelligence is the effort to create computers capable of intelligent behavior. It is a broad catchall term, used to refer to everything from Siri to IBM’s Watson to powerful technologies we have yet to invent.

Some researchers distinguish between “narrow AI” — computer systems that are better than humans in some specific, well-defined field, like playing chess or generating images or diagnosing cancer — and “general AI,” systems that can surpass human capabilities in many domains. We don’t have general AI yet, but we’re starting to get a better sense of the challenges it will pose.

Narrow AI has seen extraordinary progress over the past few years. AI systems have improved dramatically at translation, at games like chess and Go, at important research biology questions like predicting how proteins fold, and at generating images. AI systems determine what you’ll see in a Google search or in your Facebook Newsfeed. They are being developed to improve drone targeting and detect missiles.

But narrow AI is getting less narrow. Once, we made progress in AI by painstakingly teaching computer systems specific concepts. To do computer vision — allowing a computer to identify things in pictures and video — researchers wrote algorithms for detecting edges. To play chess, they programmed in heuristics about chess. To do natural language processing (speech recognition, transcription, translation, etc.), they drew on the field of linguistics.

But recently, we've gotten better at creating computer systems that have generalized learning capabilities. Instead of mathematically describing detailed features of a problem, we let the computer system learn that by itself. While once we treated computer vision as a completely different problem from natural language processing or platform game playing, now we can solve all three problems with the same approaches.

Our AI progress so far has enabled enormous advances — and has also raised urgent ethical questions. When you train a computer system to predict which convicted felons will reoffend, you're using inputs from a criminal justice system biased against black people and low-income people — and so its outputs will likely be biased against black and low-income people too. Making websites more addictive can be great for your revenue but bad for your users.

Rosie Campbell at UC Berkeley's Center for Human-Compatible AI argues that these are examples, writ small, of the big worry experts have about general AI in the future. The difficulties we're wrestling with today with narrow AI don't come from the systems turning on us or wanting revenge or considering us inferior. Rather, they come from the disconnect between what we tell our systems to do and what we actually want them to do.

For example, we tell a system to run up a high score in a video game. We want it to play the game fairly and learn game skills — but if it instead has the chance to directly hack the scoring system, it will do that. It's doing great by the metric we gave it. But we aren't getting what we wanted.

In other words, our problems come from the systems being really good at achieving the goal they learned to pursue; it's just that the goal they learned in their training environment isn't the outcome we actually wanted. And we're building systems we don't understand, which means we can't always anticipate their behavior.

Right now the harm is limited because the systems are so limited. But it's a pattern that could have even graver consequences for human beings in the future as AI systems become more advanced.

2) Is it even possible to make a computer as smart as a person?

Yes, though current AI systems aren't nearly that smart.

One popular adage about AI is "everything that's easy is hard, and everything that's hard is easy." Doing complex calculations in the blink of an eye? Easy. Looking at a picture and telling you whether it's a dog? Hard (until very recently).

Lots of things humans do are still outside AI's grasp. For instance, it's hard to design an AI system that explores an unfamiliar environment, that can navigate its way from, say, the entryway of a building it's never been in before up the stairs to a specific person's desk. We don't know how to design an AI system that reads a book and retains an understanding of the concepts.

The paradigm that has driven many of the biggest breakthroughs in AI recently is called “deep learning.” Deep learning systems can do some astonishing stuff: beat games we thought humans might never lose, invent compelling and realistic photographs, solve open problems in molecular biology.

These breakthroughs have made some researchers conclude it’s time to start thinking about the dangers of more powerful systems, but skeptics remain. The field’s pessimists argue that programs still need an extraordinary pool of structured data to learn from, require carefully chosen parameters, or work only in environments designed to avoid the problems we don’t yet know how to solve. They point to self-driving cars, which are still mediocre under the best conditions despite the billions that have been poured into making them work.

With all those limitations, one might conclude that even if it’s possible to make a computer as smart as a person, it’s certainly a long way away. But that conclusion doesn’t necessarily follow.

That’s because for almost all the history of AI, we’ve been held back in large part by not having enough computing power to realize our ideas fully. Many of the breakthroughs of recent years — AI systems that learned how to play Atari games, generate fake photos of celebrities, fold proteins, and compete in massive multiplayer online strategy games — have happened because that’s no longer true. Lots of algorithms that seemed not to work at all turned out to work quite well once we could run them with more computing power.

And the cost of a unit of computing time keeps falling. Progress in computing speed has slowed recently, but the cost of computing power is still estimated to be falling by a factor of 10 every 10 years. Through most of its history, AI has had access to less computing power than the human brain. That’s changing. By most estimates, we’re now approaching the era when AI systems can have the computing resources that we humans enjoy.

Furthermore, breakthroughs in a field can often surprise even other researchers in the field. “Some have argued that there is no conceivable risk to humanity [from AI] for centuries to come,” wrote UC Berkeley professor Stuart Russell, “perhaps forgetting that the interval of time between Rutherford’s confident assertion that atomic energy would never be feasibly extracted and Szilárd’s invention of the neutron-induced nuclear chain reaction was less than twenty-four hours.”

There’s another consideration. Imagine an AI that is inferior to humans at everything, with one exception: It’s a competent engineer that can build AI systems very effectively. Machine learning engineers who work on automating jobs in other fields often observe, humorously, that in some respects, their own field looks like one where much of the work — the tedious tuning of parameters — could be automated.

If we can design such a system, then we can use its result — a better engineering AI — to build another, even better AI. This is the mind-bending scenario experts call “recursive self-improvement,” where gains in AI capabilities enable more gains in AI capabilities, allowing a system that started out behind us to rapidly end up with abilities well beyond what we anticipated.

This is a possibility that has been anticipated since the first computers. I.J. Good, a colleague of Alan Turing who worked at the Bletchley Park codebreaking operation during World War II and helped build the first computers afterward, may have been the first to spell it out, back in 1965: "An ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

3) How exactly could it wipe us out?

It's immediately clear how nuclear bombs will kill us. No one working on mitigating nuclear risk has to start by explaining why it'd be a bad thing if we had a nuclear war.

The case that AI could pose an existential risk to humanity is more complicated and harder to grasp. So many of the people who are working to build safe AI systems have to start by explaining why AI systems, by default, are dangerous.

Javier Zarracina/Vox

The idea that AI can become a danger is rooted in the fact that AI systems pursue their goals, whether or not those goals are what we really intended — and whether or not we're in the way. "You're probably not an evil ant-hater who steps on ants out of malice," Stephen Hawking wrote, "but if you're in charge of a hydroelectric green-energy project and there's an anthill in the region to be flooded, too bad for the ants. Let's not place humanity in the position of those ants."

Here's one scenario that keeps experts up at night: We develop a sophisticated AI system with the goal of, say, estimating some number with high confidence. The AI realizes it can achieve more confidence in its calculation if it uses all the world's computing hardware, and it realizes that releasing a biological superweapon to wipe out humanity would allow it free use of all the hardware. Having exterminated humanity, it then calculates the number with higher confidence.

Victoria Krakovna, an AI researcher at DeepMind (now a division of Alphabet, Google's parent company), compiled a list of examples of "specification gaming": the computer doing what we told it to do but not what we wanted it to do. For example, we tried to teach AI organisms in a simulation to jump, but we did it by teaching them to measure how far their "feet" rose above the ground. Instead of jumping, they learned to grow into tall vertical poles and do flips — they excelled at what we were measuring, but they didn't do what we wanted them to do.

An AI playing the Atari exploration game *Montezuma's Revenge* found a bug that let it force a key in the game to reappear, thereby allowing it to earn a higher score by exploiting the glitch. An AI playing a different game realized it could get more points by falsely inserting its name as the owner of high-value items.

Sometimes, the researchers didn't even know how their AI system cheated: "the agent discovers an in-game bug. ... For a reason unknown to us, the game does not advance to the second round but the platforms start to blink and the agent quickly gains a huge amount of points (close to 1 million for our episode time limit)."

What these examples make clear is that in any system that might have bugs or unintended behavior or behavior humans don't fully understand, a sufficiently powerful AI system might act unpredictably — pursuing its goals through an avenue that isn't the one we expected.

In his 2009 paper “The Basic AI Drives,” Steve Omohundro, who has worked as a computer science professor at the University of Illinois Urbana-Champaign and as the president of Possibility Research, argues that almost any AI system will predictably try to accumulate more resources, become more efficient, and resist being turned off or modified: “These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems.”

His argument goes like this: Because AIs have goals, they’ll be motivated to take actions that they can predict will advance their goals. An AI playing a chess game will be motivated to take an opponent’s piece and advance the board to a state that looks more winnable.

But the same AI, if it sees a way to improve its own chess evaluation algorithm so it can evaluate potential moves faster, will do that too, for the same reason: It’s just another step that advances its goal.

If the AI sees a way to harness more computing power so it can consider more moves in the time available, it will do that. And if the AI detects that someone is trying to turn off its computer mid-game, and it has a way to disrupt that, it’ll do it. It’s not that we would instruct the AI to do things like that; it’s that whatever goal a system has, actions like these will often be part of the best path to achieve that goal.

That means that any goal, even innocuous ones like playing chess or generating advertisements that get lots of clicks online, could produce unintended results if the agent pursuing it has enough intelligence and optimization power to identify weird, unexpected routes to achieve its goals.

Goal-driven systems won’t wake up one day with hostility to humans lurking in their hearts. But they will take actions that they predict will help them achieve their goal — even if we’d find those actions problematic, even horrifying. They’ll work to preserve themselves, accumulate more resources, and become more efficient. They already do that, but it takes the form of weird glitches in games. As they grow more sophisticated, scientists like Omohundro predict more adversarial behavior.

4) When did scientists first start worrying about AI risk?

Scientists have been thinking about the potential of artificial intelligence since the early days of computers. In the famous paper where he put forth the Turing test for determining if an artificial system is truly “intelligent,” Alan Turing wrote:

> Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. … There would be plenty to do in trying to keep one’s intelligence up to the standards set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control.

I.J. Good worked closely with Turing and reached the same conclusions, according to his assistant, Leslie Pendleton. In an excerpt from unpublished notes Good wrote shortly before he died in 2009, he writes about himself in third person and notes a disagreement with his younger self — while as a younger man, he thought powerful AIs might be helpful to us, the older Good expected AI to annihilate us.

> [The paper] "Speculations Concerning the First Ultra-intelligent Machine" (1965) ... began: "The survival of man depends on the early construction of an ultra-intelligent machine." Those were his words during the Cold War, and he now suspects that "survival" should be replaced by "extinction." He thinks that, because of international competition, we cannot prevent the machines from taking over. He thinks we are lemmings. He said also that "probably Man will construct the deus ex machina in his own image."

In the 21st century, with computers quickly establishing themselves as a transformative force in our world, younger researchers started expressing similar worries.

Nick Bostrom is a professor at the University of Oxford, the director of the Future of Humanity Institute, and the director of the Governance of Artificial Intelligence Program. He researches risks to humanity, both in the abstract — asking questions like why we seem to be alone in the universe — and in concrete terms, analyzing the technological advances on the table and whether they endanger us. AI, he concluded, endangers us.

In 2014, he wrote a book explaining the risks AI poses and the necessity of getting it right the first time, concluding, "once unfriendly superintelligence exists, it would prevent us from replacing it or changing its preferences. Our fate would be sealed."

Across the world, others have reached the same conclusion. Bostrom co-authored a paper on the ethics of artificial intelligence with Eliezer Yudkowsky, founder of and research fellow at the Berkeley Machine Intelligence Research Institute (MIRI), an organization that works on better formal characterizations of the AI safety problem.

Yudkowsky started his career in AI by worriedly poking holes in others' proposals for how to make AI systems safe, and has spent most of it working to persuade his peers that AI systems will, by default, be unaligned with human values (not necessarily opposed to but indifferent to human morality) — and that it'll be a challenging technical problem to prevent that outcome.

Increasingly, researchers realized that there'd be challenges that hadn't been present with AI systems when they were simple. "'Side effects' are much more likely to occur in a complex environment, and an agent may need to be quite sophisticated to hack its reward function in a dangerous way. This may explain why these problems have received so little study in the past, while also suggesting their importance in the future," concluded a 2016 research paper on problems in AI safety.

Bostrom's book *Superintelligence* was compelling to many people, but there were skeptics. "No, experts don't think superintelligent AI is a threat to humanity," argued an op-ed by Oren Etzioni, a professor of computer science at the University of Washington and CEO of the Allan Institute for Artificial Intelligence. "Yes, we are worried about the existential risk of artificial intelligence," replied a dueling op-ed by Stuart Russell, an AI pioneer and UC Berkeley professor, and Allan DaFoe, a senior research fellow at Oxford and director of the Governance of AI program there.

It's tempting to conclude that there's a pitched battle between AI-risk skeptics and AI-risk believers. In reality, they might not disagree as profoundly as you would think.

Facebook's chief AI scientist Yann LeCun, for example, is a vocal voice on the skeptical side. But while he argues we shouldn't fear AI, he still believes we ought to have people working on, and thinking about, AI safety. "Even if the risk of an A.I. uprising is very unlikely and very far in the future, we still need to think about it, design precautionary measures, and establish guidelines," he writes.

That's not to say there's an expert consensus here — far from it. There is substantial disagreement about which approaches seem likeliest to bring us to general AI, which approaches seem likeliest to bring us to *safe* general AI, and how soon we need to worry about any of this.

Many experts are wary that others are overselling their field, and dooming it when the hype runs out. But that disagreement shouldn't obscure a growing common ground; these are possibilities worth thinking about, investing in, and researching, so we have guidelines when the moment comes that they're needed.

5) Why couldn't we just shut off a computer if it got too powerful?

A smart AI could predict that we'd want to turn it off if it made us nervous. So it would try hard not to make us nervous, because doing so wouldn't help it accomplish its goals. If asked what its intentions are, or what it's working on, it would attempt to evaluate which responses are least likely to get it shut off, and answer with those. If it wasn't competent enough to do that, it might pretend to be even dumber than it was — anticipating that researchers would give it more time, computing resources, and training data.

So we might not know when it's the right moment to shut off a computer.

We also might do things that make it impossible to shut off the computer later, even if we realize eventually that it's a good idea. For example, many AI systems could have access to the internet, which is a rich source of training data and which they'd need if they're to make money for their creators (for example, on the stock market, where more than half of trading is done by fast-reacting AI algorithms).

But with internet access, an AI could email copies of itself somewhere where they'll be downloaded and read, or hack vulnerable systems elsewhere. Shutting off any one computer wouldn't help.

In that case, isn't it a terrible idea to let any AI system — even one which doesn't seem powerful enough to be dangerous — have access to the internet? Probably. But that doesn't mean it won't continue to happen.

So far, we've mostly talked about the technical challenges of AI. But from here forward, it's necessary to veer more into the politics. Since AI systems enable incredible things, there will be lots of different actors working on such systems.

There will likely be startups, established tech companies like Google (Alphabet's recently acquired startup DeepMind is frequently mentioned as an AI frontrunner), and nonprofits (the Elon Musk-founded OpenAI is another major player in the field).

There will be governments — Russia's Vladimir Putin has expressed an interest in AI, and China has made big investments. Some of them will presumably be cautious and employ safety measures, including keeping their AI off the internet. But in a scenario like this one, we’re at the mercy of the least cautious actor, whoever they may be.

That’s part of what makes AI hard: Even if we know how to take appropriate precautions (and right now we don’t), we also need to figure out how to ensure that all would-be AI programmers are motivated to take those precautions and have the tools to implement them correctly.

6) What are we doing right now to avoid an AI apocalypse?

“It could be said that public policy on AGI [artificial general intelligence] does not exist,” concluded a paper this year reviewing the state of the field.

The truth is that technical work on promising approaches is getting done, but there’s shockingly little in the way of policy planning, international collaboration, or public-private partnerships. In fact, much of the work is being done by only a handful of organizations, and it has been estimated that around 50 people in the world work full time on technical AI safety.

Bostrom’s Future of Humanity Institute has published a research agenda for AI governance: the study of “devising global norms, policies, and institutions to best ensure the beneficial development and use of advanced AI.” It has published research on the risk of malicious uses of AI, on the context of China’s AI strategy, and on artificial intelligence and international security.

The longest-established organization working on technical AI safety is the Machine Intelligence Research Institute (MIRI), which prioritizes research into designing highly reliable agents — artificial intelligence programs whose behavior we can predict well enough to be confident they’re safe. (Disclosure: MIRI is a nonprofit and I donated to its work in 2017 and 2018.)

The Elon Musk-founded OpenAI is a very new organization, less than three years old. But researchers there are active contributors to both AI safety and AI capabilities research. A research agenda in 2016 spelled out “concrete open technical problems relating to accident prevention in machine learning systems,” and researchers have since advanced some approaches to safe AI systems.

Alphabet’s DeepMind, a leader in this field, has a safety team and a technical research agenda outlined here. “Our intention is to ensure that AI systems of the future are not just ‘hopefully safe’ but robustly, verifiably safe,” it concludes, outlining an approach with an emphasis on specification (designing goals well), robustness (designing systems that perform within safe limits under volatile conditions), and assurance (monitoring systems and understanding what they’re doing).

There are also lots of people working on more present-day AI ethics problems: algorithmic bias, robustness of modern machine-learning algorithms to small changes, and transparency and interpretability of neural nets, to name just a few. Some of that research could potentially be valuable for preventing destructive scenarios.

But on the whole, the state of the field is a little bit as if almost all climate change researchers were focused on managing the droughts, wildfires, and famines we’re already facing today, with only a tiny skeleton team dedicating to forecasting the future and 50 or so researchers who work full time on coming up with a plan to turn things around.

Not every organization with a major AI department has a safety team at all, and some of them have safety teams focused only on algorithmic fairness and not on the risks from advanced systems. The US government doesn't have a department for AI.

The field still has lots of open questions — many of which might make AI look much more scary, or much less so — which no one has dug into in depth.

7) Is this really likelier to kill us all than, say, climate change?

It sometimes seems like we're facing dangers from all angles in the 21st century. Both climate change and future AI developments are likely to be transformative forces acting on our world.

Our predictions about climate change are more confident, both for better and for worse. We have a clearer understanding of the risks the planet will face, and we can estimate the costs to human civilization. They are projected to be enormous, risking potentially hundreds of millions of lives. The ones who will suffer most will be low-income people in developing countries; the wealthy will find it easier to adapt. We also have a clearer understanding of the policies we need to enact to address climate change than we do with AI.

There's intense disagreement in the field on timelines for critical advances in AI. While AI safety experts agree on many features of the safety problem, they're still making the case to research teams in their own field, and they disagree on some of the details. There's substantial disagreement on how badly it could go, and on how likely it is to go badly. There are only a few people who work full time on AI forecasting. One of the things current researchers are trying to nail down is their models and the reasons for the remaining disagreements about what safe approaches will look like.

Most experts in the AI field think it poses a much larger risk of total human extinction than climate change, since analysts of existential risks to humanity think that climate change, while catastrophic, is unlikely to lead to human extinction. But many others primarily emphasize our uncertainty — and emphasize that when we're working rapidly toward powerful technology about which there are still many unanswered questions, the smart step is to start the research now.

8) Is there a possibility that AI can be benevolent?

AI safety researchers emphasize that we shouldn't assume AI systems will be benevolent by default. They'll have the goals that their training environment set them up for, and no doubt this will fail to encapsulate the whole of human values.

When the AI gets smarter, might it figure out morality by itself? Again, researchers emphasize that it won't. It's not really a matter of "figuring out" — the AI will understand just fine that humans actually value love and fulfillment and happiness, and not just the number associated with Google on the New York Stock Exchange. But the AI's values will be built around whatever goal system it was initially built around, which means it won't suddenly become aligned with human values if it wasn't designed that way to start with.

Of course, we can build AI systems that are aligned with human values, or at least that humans can safely work with. That is ultimately what almost every organization with an artificial general intelligence division is trying to do. A success with AI could give us access to decades or centuries of technological innovation all at once.

"If we're successful, we believe this will be one of the most important and widely beneficial scientific advances ever made," writes the introduction to Alphabet's DeepMind. "From climate change to the need for radically improved healthcare, too many problems suffer from painfully slow progress, their complexity overwhelming our ability to find solutions. With AI as a multiplier for human ingenuity, those solutions will come into reach."

So, yes, AI can share our values — and transform our world for the good. We just need to solve a very hard engineering problem first.

9) I just really want to know: how worried should we be?

To people who think the worrying is premature and the risks overblown, AI safety is competing with other priorities that sound, well, a bit less sci-fi — and it's not clear why AI should take precedence. To people who think the risks described are real and substantial, it's outrageous that we're dedicating so few resources to working on them.

While machine-learning researchers are right to be wary of hype, it's also hard to avoid the fact that they're accomplishing some impressive, surprising things using very generalizable techniques, and that it doesn't seem that all the low-hanging fruit has been picked.

At a major conference in early December, Google's DeepMind cracked open a longstanding problem in biology: predicting how proteins fold. "Even though there's a lot more work to do before we're able to have a quantifiable impact on treating diseases, managing the environment, and more, we know the potential is enormous," its announcement concludes.

AI looks increasingly like a technology that will change the world when it arrives. Researchers across many major AI organizations tell us it will be like launching a rocket: something we have to get right before we hit "go." So it seems urgent to get to work learning rocketry. No matter whether or not humanity should be afraid, we should definitely be doing our homework.

Correction: This piece originally stated that Eliezer Yudkowsky is a "research scientist" at the Machine Intelligence Research Institute. It should've said "research fellow."

Do you ever struggle to figure out where to donate that will make the biggest impact? Or which kind of charities to support? Over 5 days, in 5 emails, we'll walk you through research and frameworks that will ***help you decide how much and where to give****, and other ways to do good.* ***Sign up for Future Perfect's new pop-up newsletter****.*