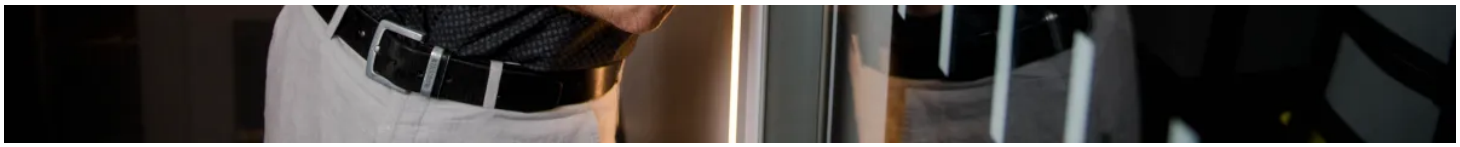
FUTURE PERFECT

He’s the godfather of AI. Now, he has a bold new plan to keep us safe from it.

“I should have thought of this 10 years ago,” Yoshua Bengio says.

by **Sigal Samuel**
Jun 19, 2025 at 6:30 AM CDT



Yoshua Bengio. AFP via Getty Images

Sigal Samuel is a senior reporter for Vox's Future Perfect and co-host of the Future Perfect podcast. She writes primarily about the future of consciousness, tracking advances in artificial intelligence and neuroscience and their staggering ethical implications. Before joining Vox, Sigal was the religion editor at the Atlantic.

The science fiction author Isaac Asimov once came up with a set of laws that we humans should program into our robots. In addition to a first, second, and third law, he also introduced a "zeroth law," which is so important that it precedes all the others: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

This month, the computer scientist Yoshua Bengio — known as the "godfather of AI" because of his pioneering work in the field — launched a new organization called LawZero. As you can probably guess, its core mission is to make sure AI won't harm humanity.

Even though he helped lay the foundation for today's advanced AI, Bengio is increasingly worried about the technology over the past few years. In 2023, he signed an open letter urging AI companies to press pause on state-of-the-art AI development. Both because of AI's present harms (like bias against marginalized

groups) and AI's future risks (like engineered bioweapons), there are very strong reasons to think that slowing down would have been a good thing.

But companies are companies. They did not slow down. In fact, they created autonomous AIs known as AI agents, which can view your computer screen, select buttons, and perform tasks — just like you can. Whereas ChatGPT needs to be prompted by a human every step of the way, an agent can accomplish multistep goals with very minimal prompting, similar to a personal assistant. Right now, those goals are simple — create a website, say — and the agents don't work that well yet. But Bengio worries that giving AIs agency is an inherently risky move: Eventually, they could escape human control and go "rogue."

So now, Bengio is pivoting to a backup plan. If he can't get companies to stop trying to build AI that matches human smarts (artificial general intelligence, or AGI) or even surpasses human smarts (artificial superintelligence, or ASI), then he wants to build something that will block those AIs from harming humanity. He calls it "Scientist AI."

Scientist AI won't be like an AI agent — it'll have no autonomy and no goals of its

own. Instead, its main job will be to calculate the probability that some other AI's action would cause harm — and, if the action is too risky, block it. AI companies could overlay Scientist AI onto their models to stop them from doing something dangerous, akin to how we put guardrails along highways to stop cars from veering off course.

I talked to Bengio about why he's so disturbed by today's AI systems, whether he regrets doing the research that led to their creation, and whether he thinks throwing yet more AI at the problem will be enough to solve it. A transcript of our unusually candid conversation, edited for length and clarity, follows.

Sigal Samuel

When people express worry about AI, they often express it as a worry about artificial general intelligence or superintelligence. Do you think that's the wrong thing to be worrying about? Should we only worry about AGI or ASI insofar as it includes agency?

Yoshua Bengio

Yes. You could have a superintelligent AI that doesn't "want" anything, and it's totally not dangerous because it doesn't have its own goals. It's just like a very smart encyclopedia.

Sigal Samuel

Researchers have been warning for years about the risks of AI systems, especially systems with their own goals and general intelligence. Can you explain what's making the situation increasingly scary to you now?

Yoshua Bengio

In the last six months, we've gotten evidence of AIs that are so misaligned that they would go against our moral instructions. They would plan and do these bad things — lying, cheating, trying to persuade us with deceptions, and — worst of all — trying to escape our control and not wanting to be shut down, and doing anything [to avoid shutdown], including blackmail. These are not an immediate danger because they're all controlled experiments…but we don't know how to really deal with this.

Sigal Samuel

And these bad behaviors increase the more agency the AI system has?

Yoshua Bengio

Yes. The systems we had last year, before we got into reasoning models, were much less prone to this. It's just getting worse and worse. That makes sense because we see that their planning ability is improving exponentially. And [the AIs] need good planning to strategize about things like "How am I going to convince these people to do what I want?" or "How do I escape their control?" So if we don't fix these problems quickly, we may end up with, initially, funny accidents, and later, not-funny accidents.

That's motivating what we're trying to do at LawZero. We're trying to think about how we design AI more precisely, so that, by construction, it's not even going to have any incentive or reason to do such things. In fact, it's not going to want anything.

Sigal Samuel

Tell me about how Scientist AI could be used as a guardrail against the bad actions of an AI agent. I'm imagining Scientist AI as the babysitter of the agentic AI, double-checking what it's doing.

Yoshua Bengio

So, in order to do the job of a guardrail, you don't need to be an agent yourself. The only thing you need to do is make a good prediction. And the prediction is this: Is this action that my agent wants to do acceptable, morally speaking? Does it satisfy the safety specifications that humans have provided? Or is it going to harm somebody? And if the answer is yes, with some probability that's not very small, then the guardrail says: No, this is a bad action. And the agent has to [try a different] action.

Sigal Samuel

But even if we build Scientist AI, the domain of "What is moral or immoral?" is famously contentious. There's just no consensus. So how would Scientist AI learn what to classify as a bad action?

Yoshua Bengio

It's not for any kind of AI to decide what is right or wrong. We should establish that using democracy. Law should be about trying to be clear about what is acceptable or not.

Now, of course, there could be ambiguity in the law. Hence you can get a corporate lawyer who is able to find loopholes in the law. But there's a way around this: Scientist AI is planned so that it will see the ambiguity. It will see that there are different interpretations, say, of a particular rule. And then it can be conservative about the interpretation — as in, if any of the plausible interpretations would judge this action as really bad, then the action is rejected.

RELATED: You can't optimize your way to being a good person

Sigal Samuel

I think a problem there would be that almost any moral choice arguably has ambiguity. We've got some of the most contentious moral issues — think about gun control or abortion in the US — where, even democratically, you might get a significant proportion of the population that says they're opposed. How do you propose to deal with that?

Yoshua Bengio

I don't. Except by having the strongest possible honesty and rationality in the answers, which, in my opinion, would already be a big gain compared to the sort of democratic discussions that are happening. One of the features of the Scientist AI, like a good human scientist, is that you can ask: Why are you saying this? And he would come up with — not "he," sorry! — *it* would come up with a justification.

The AI would be involved in the dialogue to try to help us rationalize what are the

pros and cons and so on. So I actually think that these sorts of machines could be turned into tools to help democratic debates. It's a little bit more than fact-checking — it's also like reasoning-checking.

Sigal Samuel

This idea of developing Scientist AI stems from your disillusionment with the AI we've been developing so far. And your research was very foundational in laying the groundwork for that kind of AI. On a personal level, do you feel some sense of inner conflict or regret about having done the research that laid that groundwork?

Yoshua Bengio

I should have thought of this 10 years ago. In fact, I could have, because I read some of the early works in AI safety. But I think there are very strong psychological defenses that I had, and that most of the AI researchers have. You want to feel good about your work, and you want to feel like you're the good guy, not doing something that could cause in the future lots of harm and death. So we kind of look the other way.

And for myself, I was thinking: This is so far into the future! Before we get to the science-fiction-sounding things, we're going to have AI that can help us with medicine and climate and education, and it's going to be great. So let's worry about these things when we get there.

But that was before ChatGPT came. When ChatGPT came, I couldn't continue living with this internal lie, because, well, we are getting very close to human-level.

Sigal Samuel

The reason I ask this is because it struck me when reading your plan for Scientist AI that you say it's modeled after the platonic idea of a scientist — a selfless, ideal person who's just trying to understand the world. I thought: *Are you in some way trying to build the ideal version of yourself, this "he" that you mentioned, the ideal scientist? Is it like what you wish you could have been?*

Yoshua Bengio

You should do psychotherapy instead of journalism! Yeah, you're pretty close to the

mark. In a way, it's an ideal that I have been looking toward for myself. I think that's an ideal that scientists should be looking toward as a model. Because, for the most part in science, we need to step back from our emotions so that we avoid biases and preconceived ideas and ego.

RELATED: AI systems could become conscious. What if they hate their lives?

Sigal Samuel

A couple of years ago you were one of the signatories of the letter urging AI companies to pause cutting-edge work. Obviously, the pause did not happen. For me, one of the takeaways from that moment was that we're at a point where this is not predominantly a technological problem. It's political. It's really about power and who gets the power to shape the incentive structure.

We know the incentives in the AI industry are horribly misaligned. There's massive commercial pressure to build cutting-edge AI. To do that, you need a ton of compute so you need billions of dollars, so you're practically forced to get in bed with a Microsoft or an Amazon. How do you propose to avoid that fate?

Yoshua Bengio

That's why we're doing this as a nonprofit. We want to avoid the market pressure that would force us into the capability race and, instead, focus on the scientific aspects of safety.

I think we could do a lot of good without having to train frontier models ourselves. If we come up with a methodology for training AI that is convincingly safer, at least on some aspects like loss of control, and we hand it over almost for free to companies that are building AI — well, no one in these companies actually wants to see a rogue AI. It's just that they don't have the incentive to do the work! So I think just knowing how to fix the problem would reduce the risks considerably.

I also think that governments will hopefully take these questions more and more seriously. I know right now it doesn't look like it, but when we start seeing more evidence of the kind we've seen in the last six months, but stronger and more scary, public opinion might push sufficiently that we'll see regulation or some way to incentivize companies to behave better. It might even happen just for market reasons — like, [AI companies] could be sued. So, at some point, they might reason that they should be willing to pay some money to reduce the risks of accidents.

Sigal Samuel

I was happy to see that LawZero isn't only talking about reducing the risks of accidents but is also talking about "protecting human joy and endeavor." A lot of people fear that if AI gets better than them at things, well, what is the meaning of their life? How would you advise people to think about the meaning of their human life if we enter an era where machines have both agency and extreme intelligence?

Yoshua Bengio

I understand it would be easy to be discouraged and to feel powerless. But the decisions that human beings are going to make in the coming years as AI becomes more powerful — these decisions are incredibly consequential. So there's a sense in which it's hard to get more meaning than that! If you want to do something about it, be part of the thinking, be part of the democratic debate.

I would advise us all to remind ourselves that we have agency. And we have an amazing task in front of us: to shape the future.

SEE MORE: FUTURE PERFECT LIVING IN AN AI WORLD TECHNOLOGY

MORE IN THIS STREAM SEE ALL

★ THE HIGHLIGHT

AI systems could become conscious. What if they hate their lives?

By SIGAL SAMUEL

★ THE HIGHLIGHT

My students think it’s fine to cheat with AI. Maybe they’re onto something.

By SIGAL SAMUEL

Why Pope Leo has so much to say about AI, briefly explained

By SIGAL SAMUEL

Read More

00:00 02:00

More in Future Perfect

FUTURE PERFECT | JUN 18

The one thing the Trump administration got very right

Too bad it's now sabotaging it.

By MARINA BOLOTNIKOVA

FUTURE PERFECT | JUN 15

The stunning reversal of humanity's oldest bias

Everyone wants to be a girl dad now.

By BRYAN WALSH

FUTURE PERFECT | JUN 13

What drove the tech right's — and Elon Musk's — big, failed bet on Trump

The tech right saw in Trump what they wanted to see, but it wasn't actually there

By KELSEY PIPER

FUTURE PERFECT | JUN 12

The one drug RFK Jr. should actually ban

The world's "most controversial" food additive, explained.

By KENNY TORRELLA

TECHNOLOGY | JUN 12

Your iPhone is about to get uglier

And that's not even Apple's biggest problem right now

By ADAM CLARK ESTES

FUTURE PERFECT | JUN 11

The right refuses to take AI seriously

Republicans are gutting the safety net as job-killing mass automation looms.

By DYLAN MATTHEWS

Recommended For You

JUN 18

The one thing the Trump administration got very right

JUN 18

This veteran health official watched Americans lose trust in science. How do we get it back?

JUN 16

Scientists are dropping live mosquitoes out of drones in Hawaii. Here’s why.

JUN 16

Why we’re barely keeping track of this growing climate problem

JUN 18

How climate change will worsen hunger

JUN 18

Why are so many straight guys so bad at gossiping?