THE SHIFT

A.I.’s Prophet of Doom Wants to Shut It All Down

Eliezer Yudkowsky has spent the past 20 years warning A.I. insiders of danger. Now, he’s making his case to the public.

Listen to this article · 13:32 min Learn more

By Kevin Roose
Reporting from Berkeley, Calif.

Sept. 12, 2025

The first time I met Eliezer Yudkowsky, he said there was a 99.5 percent chance that A.I. was going to kill me.

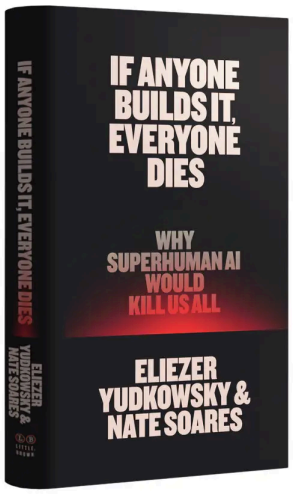
I didn’t take it personally. Mr. Yudkowsky, 46, is the founder of the Machine Intelligence Research Institute, a Berkeley-based nonprofit that studies risks from advanced artificial intelligence.

For the last two decades, he has been Silicon Valley’s version of a doomsday preacher — telling anyone who will listen that building powerful A.I. systems is a terrible idea, one that will end in disaster.

That is also the message of Mr. Yudkowsky’s new book, “If Anyone Builds It, Everyone Dies.” The book, co-written with MIRI’s president, Nate Soares, is a distilled, mass-market version of the case they have been making to A.I. insiders for years.

Their goal is to stop the development of A.I. — and the stakes, they say, are existential.

"If any company or group, anywhere on the planet, builds an artificial superintelligence using anything remotely like current techniques, based on anything remotely like the present understanding of A.I., then everyone, everywhere on Earth, will die," they write.



Mr. Yudkowsky wrote "If Anyone Builds It, Everyone Dies" with Nate Soares.

This kind of blunt doomsaying has gotten Mr. Yudkowsky dismissed by some as an extremist or a crank. But he is a central figure in modern A.I. history, and his influence on the industry is undeniable.

He was among the first people to warn of risks from powerful A.I. systems, and many A.I. leaders, including OpenAI's Sam Altman and Elon Musk, have cited his ideas. (Mr. Altman has said Mr. Yudkowsky was "critical in the decision to start OpenAI," and suggested that he might deserve a Nobel Peace Prize.)

Google, too, owes some of its A.I. success to Mr. Yudkowsky. In 2010, he introduced the founders of DeepMind — a London start-up that was trying to build advanced A.I. systems — to Peter Thiel, the venture capitalist. Mr. Thiel became DeepMind's first major investor, before Google acquired the company in 2014. Today, DeepMind's co-founder Demis Hassabis oversees Google's A.I. efforts.

Mr. Yudkowsky introduced the founders of the A.I. lab DeepMind to Peter Thiel, who became their first major financial backer. Getty Images

Demis Hassabis, a DeepMind founder, went on to run Google’s A.I. work and shared a Nobel Prize for A.I. research. Camille Cohen/Agence France-Presse — Getty Images

In addition to his work on A.I. safety — a field he more or less invented — Mr. Yudkowsky is the intellectual force behind Rationalism, a loosely organized movement (or a religion, depending on whom you ask) that pursues self-improvement through rigorous reasoning. Today, Silicon Valley tech companies are full of young Rationalists, many of whom grew up reading Mr. Yudkowsky’s writing online.

I’m not a Rationalist, and my view of A.I. is considerably more moderate than Mr. Yudkowsky’s. (I don’t, for instance, think we should bomb data centers if rogue nations threaten to develop superhuman A.I. in violation of international agreements, a view he has espoused.) But in recent months, I’ve sat down with him several times to better understand his views.

At first, he resisted being profiled. (Ideas, not personalities, are what he thinks rational people should care about.) Eventually, he agreed, in part because he hopes that by sharing his fears about A.I., he might persuade others to join the cause of saving humanity.

“To have the world turn back from superintelligent A.I., and we get to not die in the immediate future,” he told me. “That’s all I presently want out of life.”

From 'Friendly A.I.' to 'Death With Dignity'

Mr. Yudkowsky grew up in an Orthodox Jewish family in Chicago. He dropped out of school after eighth grade because of chronic health issues, and never returned. Instead, he devoured science fiction books, taught himself computer science and started hanging out online with a group of far-out futurists known as the Extropians.

He was enchanted by the idea of the singularity — a hypothetical future point when A.I. would surpass human intelligence. And he wanted to build an artificial general intelligence, or A.G.I., an A.I. system capable of doing everything the human brain can.

"He seemed to think A.G.I. was coming soon," said Ben Goertzel, an A.I. researcher who met Mr. Yudkowsky as a teenager. "He also seemed to think he was the only person on the planet smart enough to create A.G.I."

He moved to the Bay Area in 2005 to pursue what he called "friendly A.I." — A.I. that would be aligned with human values, and would care about human well-being.

But the more Mr. Yudkowsky learned, the more he came to believe that building friendly A.I. would be difficult, if not impossible.

One reason is what he calls "orthogonality" — the notion that intelligence and benevolence are separate traits, and that an A.I. system would not automatically get friendlier as it got smarter.

Another is what he calls "instrumental convergence" — the idea that a powerful, goal-directed A.I. system could adopt strategies that end up harming humans. (A well-known example is the "paper clip maximizer," a thought experiment popularized by the philosopher Nick Bostrom, based on what Mr. Yudkowsky claims is a misunderstanding of an idea of his, in which an A.I. is told to maximize paper clip production and destroys humanity to gather more raw materials.)

He also worried about what he called an "intelligence explosion" — a sudden, drastic spike in A.I. capabilities that could lead to the rapid emergence of superintelligence.

At the time, these were abstract, theoretical arguments hashed out among internet futurists. Nothing remotely like today's A.I. systems existed, and the idea of a rogue, superintelligent A.I. was too far-fetched for serious scientists to worry about.

Kevin Roose and Casey Newton are the hosts of Hard Fork, a podcast that makes sense of the rapidly changing world of technology. Subscribe and listen.

But over time, as A.I. capabilities improved, Mr. Yudkowsky's ideas found a wider audience.

In 2010, he started writing "Harry Potter and the Methods of Rationality," a serialized work of Harry Potter fan fiction that he hoped would introduce more people to the core concepts of Rationalism. The book eventually sprawled to more than 600,000 words — longer than "War and Peace." (Brevity is not Mr. Yudkowsky's strong suit — another of his works, a B.D.S.M.-themed Dungeons & Dragons fan fiction that contains his views of decision theory, clocks in at 1.8 million words.)

Despite its length, "Harry Potter and the Methods of Rationality" was a cult hit, and introduced legions of young people to Mr. Yudkowsky's worldview. Even today, I routinely meet employees of top A.I. companies who tell me, somewhat sheepishly, that reading the book inspired their career choice.

Mr. Yudkowsky's writing has been the foundation of the influential Rationalist movement. Loren Elliott for The New York Times

Some young Rationalists went to work for MIRI, Mr. Yudkowsky's organization. Others fanned out across the tech industry, taking jobs at companies like OpenAI and Google. But nothing they did slowed the pace of A.I. progress, or allayed any of Mr. Yudkowsky's fears about how powerful A.I. would turn out.

In 2022, Mr. Yudkowsky announced — in what some interpreted as an April Fools joke — that he and MIRI were pivoting to a new strategy he called "death with dignity." Humanity was doomed to die, he said, and instead of continuing to fight a losing battle to align A.I. with human values, he was shifting his focus to helping people accept their fate.

"It's obvious at this point that humanity isn't going to solve the alignment problem, or even try very hard, or even go out with much of a fight," he wrote.

Are We Really Doomed?

These are, it should be said, extreme views even by the standards of A.I. pessimists. And during our most recent conversation, I raised some objections to Mr. Yudkowsky's claims.

Haven't researchers made strides in areas like mechanistic interpretability — the field that studies the inner workings of A.I. models — that may give us better ways of controlling powerful A.I. systems?

"The course of events over the last 25 years has not been such as to invalidate any of these underlying theories," he replied. "Imagine going up to a physicist and saying, 'Have any of the recent discoveries in physics changed your mind about rocks falling off cliffs?'"

What about the more immediate harms that A.I. poses — such as job loss, and people falling into delusional spirals while talking to chatbots? Shouldn't he be at least as focused on those as on doomsday scenarios?

Mr. Yudkowsky acknowledged that some of these harms were real, but scoffed at the idea that he should focus more on them.

"It's like saying that Leo Szilard, the person who first conceived of the nuclear chain reactions behind nuclear weapons, ought to have spent all of his time and energy worrying about the current harms of the Radium Girls," a group of young women who developed radiation poisoning while painting watch dials in factories in the 1920s.

Are there any A.I. companies he's rooting for? Any approaches less likely to lead us to doom?

"Among the crazed mad scientists driving headlong toward disaster, every last one of which should be shut down, OpenAI's management is noticeably worse than the pack, and some of Anthropic's employees are noticeably better than the pack," he said. "None of this makes a difference, and all of them should be treated the same way by the law."

And what about the good things that A.I. can do? Wouldn't shutting down A.I. development also mean delaying cures for diseases, A.I. tutors for students and other benefits?

"We totally acknowledge the good effects," he replied. "Yep, these things could be great tutors. Yep, these things sure could be useful in drug discovery. Is that worth exterminating all life on Earth? No."

Does he worry about his followers committing acts of violence, going on hunger strikes or carrying out other extreme acts in order to stop A.I.?

"Humanity seems more likely to perish of doing too little here than too much," he said. "Our society refusing to have a conversation about a threat to all life on Earth, because somebody else might possibly take similar words and mash them together and then say or do something stupid, would be a foolish way for humanity to go extinct."

One Last Battle

Even among his fans, Mr. Yudkowsky is a divisive figure. He can be arrogant and abrasive, and some of his followers wish he were a more polished spokesman. He has also had to adjust to writing for a mainstream audience, rather than for Rationalists who will wade through thousands of dense, jargon-packed pages.

"He wrote 300 percent of the book," his co-author, Mr. Soares, quipped. "I wrote another negative 200 percent."

He has adjusted his image in preparation for his book tour. He shaved his beard down from its former, rabbinical length and replaced his signature golden top hat with a muted newsboy cap. (The new hat, he said dryly, was "a result of observer feedback.")

Nate Soares, right, Mr. Yudkowsky’s co-author, jokes that he trimmed down Mr. Yudkowsky’s voluminous writing. via Gretta Duleba

Several months before his new book’s release, a fan suggested he take one of his self-published works, an erotic fantasy novel, off Amazon to avoid alienating potential readers. (He did, but grumbled to me that “it’s not actually even all that horny, by my standards.”)

In 2023, he started dating Gretta Duleba, a relationship therapist, and moved to Washington State, far from the Bay Area tech bubble. To his friends, he seems happier now, and less inclined to throw in the towel on humanity’s existence.

Even the way he talks about doom has changed. He once confidently predicted, with mathematical precision, how long it would take for superhuman A.I. to be developed. But he now balks at those exercises.

“What is this obsession with timelines?” he asked. “People used to trade timelines the way that they traded astrological signs, and now they’re trading probabilities of everybody dying the way they used to trade timelines. If the probability is quite large and you don’t know when it’s going to happen, deal with it. Stop making up these numbers.”

I'm not persuaded by the more extreme parts of Mr. Yudkowsky's arguments. I don't think A.I. alignment is a lost cause, and I worry less about "Terminator"-style takeovers than about the more mundane ways A.I. could steer us toward disaster. My p(doom) — a rough, vibes-based measure of how probable I feel A.I. catastrophe is — hovers somewhere between 5 and 10 percent, making me a tame moderate by comparison.

I also think there is essentially no chance of stopping A.I. development through international treaties and nuclear-weapon-level controls on A.I. chips, as Mr. Yudkowsky and Mr. Soares argue is necessary. The Trump administration is set on accelerating A.I. progress, not slowing it down, and "doomer" has become a pejorative in Washington.

Even under a different White House administration, hundreds of millions of people would be using A.I. products like ChatGPT every day, with no clear signs of impending doom. And absent some obvious catastrophe, A.I.'s benefits would seem too obvious, and the risks too abstract, to hit the kill switch now.

But I also know that Mr. Yudkowsky and Mr. Soares have been thinking about A.I. risks far longer than most, and that there are still many reasons to worry about A.I. For starters, A.I. companies still don't really understand how large language models work, or how to control their behavior.

Their brand of doomsaying isn't popular these days. But in a world of mealy-mouthed pablum about "maximizing the benefits and minimizing the risks" of A.I., maybe they deserve some credit for putting their cards on the table.

"If we get an effective international treaty shutting A.I. down, and the book had something to do with it, I'll call the book a success," Mr. Yudkowsky told me. "Anything other than that is a sad little consolation prize on the way to death."

Kevin Roose is a Times technology columnist and a host of the podcast "Hard Fork."

A version of this article appears in print on , Section B, Page 1 of the New York edition with the headline: A.I. Prophet Wants It All Shut Down