POPULAR SCIENCE

Want More?

Can AI escape our control and destroy us?

Skype cofounder Jaan Tallinn bankrolls efforts to keep superintelligent AI under control.

By MARA HVISTENDAHL MAY 21, 2019

"It began three and a half billion years ago in a pool of muck, when a molecule made a copy of itself and so became the ultimate ancestor of all earthly life. It began four million years ago, when brain volumes began climbing rapidly in the hominid line. Fifty thousand years ago with the rise of Homo sapiens. *Ten thousand years ago with the invention of civilization. Five hundred years ago with the invention of the printing press. Fifty years ago with the invention of the computer. In less than thirty years, it will end."*

Jaan Tallinn stumbled across these words in 2007, in an online essay called "Staring into the Singularity." The "it" is human civilization. Humanity would cease to exist, predicted the essay's author, with the emergence of superintelligence, or AI that surpasses the human intellect in a broad array of areas.

Tallinn, an Estonia-born computer programmer, has a background in physics and a propensity to approach life like one big programming problem. In 2003, he had co-founded Skype, developing the backend for the app. He cashed in his shares after eBay bought it two years later, and now he was casting about for something to do. "Staring into the Singularity" mashed up computer code, quantum physics, and *Calvin and Hobbes* quotes. He was hooked.

Tallinn soon discovered that the essay's author, self-taught theorist Eliezer Yudkowsky, had written more than 1,000 articles and blog posts, many of them devoted to superintelligence. Tallinn wrote a program to scrape Yudkowsky's writings from the internet, order them chronologically, and format them for his iPhone. Then he spent the better part of a year reading them.

The term "artificial intelligence," or the simulation of intelligence in computers or machines, was coined back in 1956, only a decade after the creation of the first electronic digital computers. Hope for the field was initially high, but by the 1970s, when early predictions did not pan out, an "AI winter" set in. When Tallinn found Yudkowsky's essays, AI was undergoing a renaissance. Scientists were developing AIs that excelled in specific areas, such as winning at chess, cleaning the kitchen floor, and recognizing human speech. (In 2007, the resounding win at *Jeopardy!* of IBM's Watson was still four years away, while the triumph at Go of DeepMind's AlphaGo was eight years off.) Such "narrow" AIs, as they're called, have superhuman capabilities, but only in their specific areas of dominance. A chess-playing AI can't clean the floor or take you from point A to point B. But super-intelligent AI, Tallinn came to believe, will combine a wide range of skills in one entity. More darkly, it also might use data generated by smartphone-toting humans to excel at social manipulation.

Reading Yudkowsky's articles, Tallinn became convinced that superintelligence could lead to an explosion or "breakout" of AI that could threaten human existence—that ultrasmart AIs will take our place on the evolutionary ladder and dominate us the way we now dominate apes. Or, worse yet, exterminate us.

After finishing the last of the essays, Tallinn shot off an email to Yudkowsky—all lowercase, as is his style. "i'm jaan, one of the founding engineers of skype," he wrote. Eventually he got to the point: "i do agree that…preparing for the event of general AI surpassing human intelligence is one of the top tasks for humanity." He wanted to help. When he flew to the Bay Area for other meetings soon after, he met Yudkowsky at a Panera Bread in Millbrae, California, near where he lives. Their get-together stretched to four hours. "He actually, genuinely understood the underlying concepts and the details," Yudkowsky recalls. "This is very rare." Afterward, Tallinn wrote a check for $5,000 to the Singularity Institute for Artificial Intelligence, the nonprofit where Yudkowsky was a research fellow. (The organization changed its name to Machine Intelligence Research Institute, or MIRI, in 2013.) Tallinn has since given it more than $600,000.

The encounter with Yudkowsky brought Tallinn purpose, sending him on a mission to save us from our own creations. As he connected on the issue with other theorists and computer scientists, he embarked on a life of travel, giving talks around the world on the threat posed by superintelligence. Mostly, though, he began funding research into methods that might give humanity a way out: so-called friendly AI. That doesn't mean a machine or agent is particularly skilled at chatting about the weather, or that it remembers the names of your kids—though super-intelligent AI might be able to do both of those things. It doesn't mean it is motivated by altruism or love. A common fallacy is assuming that AI has human urges and values. "Friendly" means something much more fundamental: that the machines of tomorrow will not wipe us out in their quest to attain their goals.

Nine years after his meeting with Yudkowsky, Tallinn joins me for a meal in the dining hall of Cambridge University's Jesus College. The churchlike space is bedecked with stained-glass windows, gold molding, and oil paintings of men in wigs. Tallinn sits at a heavy mahogany table, wearing the casual garb of Silicon Valley: black jeans, T-shirt, canvas sneakers. A vaulted timber ceiling extends high above his shock of gray-blond hair.

At 46, Tallinn is in some ways your textbook tech entrepreneur. He thinks that thanks to advances in science (and provided AI doesn't destroy us), he will live for "many, many years." His concern about superintelligence is common among his cohort. PayPal co-founder Peter Thiel's foundation has given $1.6 million to MIRI, and in 2015, Tesla founder Elon Musk donated $10 million to the Future of Life Institute, a technology safety organization in Cambridge, Massachusetts. Tallinn's entrance to this rarefied world came behind the Iron Curtain in the 1980s, when a classmate's father with a government job gave a few bright kids access to mainframe computers. After Estonia became independent, he founded a video-game company. Today, Tallinn still lives in its capital city—which by a quirk of etymology is also called Tallinn—with his wife and the youngest of his six kids. When he wants to meet with researchers, he often just flies them to the Baltic region.

His giving strategy is methodical, like almost everything else he does. He spreads his money among 11 organizations, each working on different approaches to AI safety, in the hope that one might stick. In 2012, he co-founded the Cambridge Centre for the Study of Existential Risk (CSER) with an initial outlay of close to $200,000.

Existential risks—or X-risks, as Tallinn calls them—are threats to humanity's survival. In addition to AI, the 20-odd researchers at CSER study climate change, nuclear war, and bioweapons. But to Tallinn, the other disciplines mostly help legitimize the threat of runaway artificial intelligence. "Those are really just gateway drugs," he tells me. Concern about more widely accepted threats, such as climate change, might draw people in. The horror of super-intelligent machines taking over the world, he hopes, will convince them to stay. He is here now for a conference because he wants the academic community to take AI safety seriously.

Our dining companions are a random assortment of conference-goers, including a woman from Hong Kong who studies robotics and a British man who graduated from Cambridge in the 1960s. The older man asks everybody at the table where they attended university. (Tallinn's answer, Estonia's University of Tartu, does not impress him.) He

then tries to steer the conversation toward the news. Tallinn looks at him blankly. “I am not interested in near-term risks,” he says.

Tallinn changes the topic to the threat of superintelligence. When not talking to other programmers, he defaults to metaphors, and he runs through his suite of them now: Advanced AI can dispose of us as swiftly as humans chop down trees. Superintelligence is to us what we are to gorillas. Inscribed in Latin above his head is a line from Psalm 133: “How good and how pleasant it is for brothers to dwell together in unity.” But unity is far from what Tallinn has in mind in a future containing a rogue superintelligence.

An AI would need a body to take over, the older man says. Without some kind of physical casing, how could it possibly gain physical control? Tallinn has another metaphor ready: “Put me in a basement with an internet connection, and I could do a lot of damage,” he says. Then he takes a bite of risotto.

Whether a Roomba or one of its world-dominating descendants, an AI is driven by outcomes. Programmers assign these goals, along with a series of rules on how to pursue them. Advanced AI wouldn’t necessarily need to be given the goal of world domination in order to achieve it—it could just be accidental. And the history of computer programming is rife with small errors that sparked catastrophes. In 2010, for example, a trader working for the mutual-fund company Waddell & Reed sold thousands of futures contracts. The firm’s software left out a key variable from the algorithm that helped execute the trade. The result was the trillion-dollar U.S. “flash crash.”

The researchers Tallinn funds believe that if the reward structure of a superhuman AI is not properly programmed, even benign objectives could have insidious ends. One well-known example, laid out by Oxford University philosopher Nick Bostrom in his book *Superintelligence*, is a fictional agent directed to make as many paper clips as possible. The AI might decide that the atoms in human bodies would be put to better use as raw material for them.

Tallinn’s views have their share of detractors, even among the community of people concerned with AI safety. Some object that it is too early to worry about restricting super-intelligent AI when we don’t yet understand it. Others say that focusing on rogue technological actors diverts attention from the most urgent problems facing the field, like the fact that the majority of algorithms are designed by white men, or based on data biased toward them. “We’re in danger of building a world that we don’t want to live in if we don’t address those challenges in the near term,” says Terah Lyons, executive director of the Partnership on AI, a multistakeholder organization focused on AI safety and other issues. (Several of the institutes Tallinn backs are members.) But, she adds, some of the near-term challenges facing researchers—such as weeding out algorithmic bias—are precursors to ones that humanity might see with super-intelligent AI.

Tallinn isn’t so convinced. He counters that super-intelligent AI brings unique threats. Ultimately, he hopes that the AI community might follow the lead of the anti-nuclear movement in the 1940s. In the wake of the bombings of Hiroshima and Nagasaki, scientists banded together to try to limit further nuclear testing. “The Manhattan Project scientists could have said, ‘Look, we are doing innovation here, and innovation is always good, so let’s just plunge ahead,’” he tells me. “But they were more responsible than that.”

Tallinn warns that any approach to AI safety will be hard to get right. If an AI is sufficiently smart, he explains, it might have a better understanding of the constraints than its creators do. Imagine, he says, “waking up in a prison built by a bunch of blind 5-year-olds.” That is what it might be like for a super-intelligent AI that is confined by humans.

Yudkowsky, the theorist, found evidence this might be true when, starting in 2002, he conducted chat sessions in which he played the role of an AI enclosed in a box, while a rotation of other people played the gatekeeper tasked with keeping the AI in. Three out of five times, Yudkowsky—a mere mortal—says he convinced the gatekeeper to release him. His experiments have not discouraged researchers from trying to design a better box, however.

The researchers that Tallinn funds are pursuing a broad variety of strategies, from the practical to the seemingly far-fetched. Some theorize about boxing AI, either physically, by building an actual structure to contain it, or by

programming in limits to what it can do. Others are trying to teach AI to adhere to human values. A few are working on a last-ditch off switch. One researcher who is delving into all three is mathematician and philosopher Stuart Armstrong at the University of Oxford's Future of Humanity Institute, which Tallinn calls "the most interesting place in the universe." (Tallinn has given FHI more than $310,000.) Armstrong is one of the few researchers in the world who focuses full time on AI safety.

I meet him for coffee one afternoon in a cafe in Oxford. He wears a rugby shirt unbuttoned at the collar, and has the look of someone who spends his life behind a screen, with a pale face framed by a mess of sandy hair. He peppers his explanations with a disorienting mixture of popular-culture references and math. When I ask him what it might look like to succeed at AI safety, he says: "Have you seen *The Lego Movie*? Everything is awesome."

One strain of Armstrong's research looks at a specific approach to boxing called an "oracle" AI. In a 2012 paper with Nick Bostrom, who co-founded FHI, he proposed not only walling off superintelligence in a holding tank—a physical structure—but also restricting it to answering questions, like a really smart Ouija board. Even with these boundaries, an AI would have immense power to reshape the fate of humanity by subtly manipulating its interrogators. To reduce the possibility of this happening, Armstrong has proposed time limits on conversations, or banning questions that might upend the current world order. He also has suggested giving the oracle proxy measures of human survival, such as the Dow Jones Industrial Average or the number of people crossing the street in Tokyo, and telling it to keep these steady.

Ultimately, Armstrong believes, it could be necessary to create, as he calls it in one paper, a "big red off button": either a physical switch, or a mechanism programmed into an AI to automatically turn itself off in the event of a breakout. But designing such a switch is far from easy. It's not just that an advanced AI interested in self-preservation could prevent the button from being pressed. It also could become curious about why humans devised the button, activate it to see what happens, and render itself useless. In 2013, a programmer named Tom Murphy VII designed an AI that could teach itself to play Nintendo Entertainment System games. Determined not to lose at *Tetris*, the AI simply pressed pause—and kept the game frozen. "Truly, the only winning move is not to play," Murphy observed wryly, in a paper on his creation.

For the strategy to succeed, an AI has to be uninterested in the button, or, as Tallinn puts it, "it has to assign equal value to the world where it's not existing and the world where it's existing." But even if researchers can achieve that, there are other challenges. What if the AI has copied itself several thousand times across the internet?

The approach that most excites researchers is finding a way to make AI adhere to human values—not by programming them in, but by teaching AIs to learn them. In a world dominated by partisan politics, people often dwell on the ways in which our principles differ. But, Tallinn notes, humans have a lot in common: "Almost everyone values their right leg. We just don't think about it." The hope is that an AI might be taught to discern such immutable rules.

In the process, an AI would need to learn and appreciate humans' less-than-logical side: that we often say one thing and mean another, that some of our preferences conflict with others, and that people are less reliable when drunk. But the data trails we all leave in apps and social media might provide a guide. Despite the challenges, Tallinn believes, we must try because the stakes are so high. "We have to think a few steps ahead," he says. "Creating an AI that doesn't share our interests would be a horrible mistake."

On Tallinn's last night in Cambridge, I join him and two researchers for dinner at a British steakhouse. A waiter seats our group in a white-washed cellar with a cave-like atmosphere. He hands us a one-page menu that offers three different kinds of mash. A couple sits down at the table next to us, and then a few minutes later asks to move elsewhere. "It's too claustrophobic," the woman complains. I think of Tallinn's comment about the damage he could wreak if locked in a basement with nothing but an internet connection. Here we are, in the box. As if on cue, the men contemplate ways to get out.

Tallinn's guests include former genomics researcher Seán Ó hÉigeartaigh, who is CSER's executive director, and

Matthijs Maas, an AI policy researcher at the University of Copenhagen. They joke about an idea for a nerdy action flick titled *Superintelligence vs. Blockchain!*, and discuss an online game called *Universal Paperclips*, which riffs on the scenario in Bostrom's book. The exercise involves repeatedly clicking your mouse to make paper clips. It's not exactly flashy, but it does give a sense for why a machine might look for more-expedient ways to produce office supplies.

Eventually, talk shifts toward bigger questions, as it often does when Tallinn is present. The ultimate goal of AI-safety research is to create machines that are, as Cambridge philosopher and CSER co-founder Huw Price once put it, "ethically as well as cognitively superhuman." Others have raised the question: If we don't want AI to dominate us, do we want to dominate it? In other words, does AI have rights? Tallinn says this is needless anthropomorphizing. It assumes that intelligence equals consciousness—a misconception that annoys many AI researchers. Earlier in the day, CSER researcher Jose Hernandez-Orallo joked that when speaking with AI researchers, consciousness is "the C-word." ("And 'free will' is the F-word," he added.)

RELATED: What it's really like working as a safety driver in a self-driving car

---

In the cellar now, Tallinn says that consciousness is beside the point: "Take the example of a thermostat. No one would say it is conscious. But it's really inconvenient to face up against that agent if you're in a room that is set to negative 30 degrees."

Ó hÉigeartaigh chimes in. "It would be nice to worry about consciousness," he says, "but we won't have the luxury to worry about consciousness if we haven't first solved the technical safety challenges."

People get overly preoccupied with what super-intelligent AI is, Tallinn says. What form will it take? Should we worry about a single AI taking over, or an army of them? "From our perspective, the important thing is what AI does," he stresses. And that, he believes, may still be up to humans—for now.

This article was originally published in the Winter 2018 Danger issue of Popular Science.

Latest News

Want more news like this?

Sign up to receive our email newsletter and never miss an update!

Enter email address

Sign Up

By submitting above, you agree to our privacy policy.

Many products featured on this site were editorially chosen. Popular Science may receive financial compensation for products purchased through this site.