The AI Anxiety

Some Scientists Fear Superintelligent Machines Could Pose a Threat to Humanity

Joel Achenbach
December 17, 2015
The Washington Post

The world's spookiest philosopher is Nick Bostrom, a thin, soft-spoken Swede. Of all the people worried about runaway artificial intelligence, and Killer Robots, and the possibility of a technological doomsday, Bostrom conjures the most extreme scenarios. In his mind, human extinction could be just the beginning.

Bostrom's favorite apocalyptic hypothetical involves a machine that has been programmed to make paper clips (although any mundane product will do). This machine keeps getting smarter and more powerful, but never develops human values. It achieves "superintelligence." It begins to convert all kinds of ordinary materials into paper clips. Eventually it decides to turn everything on Earth — including the human race (!!!) — into paper clips.

Then it goes interstellar.

"You could have a superintelligence whose only goal is to make as many paper clips as possible, and you get this bubble of paper clips spreading through the universe," Bostrom calmly told an audience in Santa Fe, N.M., earlier this year.

He added, maintaining his tone of understatement, "I think that would be a low-value future."

The influential 42-year-old philosopher Nick Bostrom favors the creation of "superintelligent" computers, but only if done with great vigilance, with safeguards to ensure that the machines do not escape human control and pose an existential threat to humanity.

Bostrom's underlying concerns about machine intelligence, unintended consequences and potentially malevolent computers have gone mainstream. You can't attend a technology conference these days without someone bringing up the A.I. anxiety. It hovers over the tech conversation with the high-pitched whine of a 1950s-era Hollywood flying saucer.

People will tell you that even Stephen Hawking is worried about it. And Bill Gates. And that Elon Musk gave $10 million for research on how to keep machine intelligence under control. All that is true.

How this came about is as much a story about media relations as it is about technological change. The machines are not on the verge of taking over. This is a topic rife with speculation and perhaps a whiff of hysteria.

But the discussion reflects a broader truth: We live in an age in which machine intelligence has become a part of daily life. Computers fly planes and soon will drive cars. Computer algorithms anticipate our needs and decide which advertisements to show us. Machines create news stories without human intervention. Machines can recognize your face in a crowd.

New technologies — including genetic engineering and nanotechnology — are cascading upon one another and converging. We don't know how this will play out. But some of the most serious thinkers on Earth worry about potential hazards — and wonder whether we remain fully in control of our inventions.

"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."

The Singularity

Science fiction pioneer Isaac Asimov anticipated these concerns when he began writing about robots in the 1940s. He developed rules for robots, the first of which was: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

People still talk about Asimov's rules. But they talk even more about what they call "the Singularity."

The idea dates to at least 1965, when British mathematician and code-breaker I.J. Good wrote, "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."

In 1993, science fiction author Vernor Vinge used the term "the Singularity" to describe such a moment. Inventor and writer Ray Kurzweil ran with the idea, cranking out a series of books predicting the age of intelligent, spiritual machines.

Kurzweil, now a director of engineering at Google, embraces such a future; he is perhaps the most famous of the techno-utopians, for he believes that technological progress will culminate in a merger of human and machine intelligence. We will all become "transhuman."

Whether any of this actually will happen is the subject of robust debate. Bostrom supports the research but worries that sufficient safeguards are not in place.

Imagine, Bostrom says, that human engineers programmed the machines to never harm humans — an echo of the first of Asimov's robot laws. But the machines might decide that the best way to obey the harm-no-humans command would be to prevent any humans from ever being born.

Or imagine, Bostrom says, that superintelligent machines are programmed to ensure that whatever they do will make humans smile. They may then decide that they should implant electrodes into the facial muscles of all people to keep us smiling.

Bostrom isn't saying this will happen. These are thought experiments. His big-picture idea is that, just in the past couple of hundred years, we've seen astonishing changes in the human population and economic prosperity. In Bostrom's view, our modern existence is an anomaly — one created largely by technology. Our tools have suddenly overwhelmed the restrictions of nature. We're in charge now, or seem to be for the moment.

But what if the technology bites back?

“The future is ours to shape. I feel we are in a race that we need to win. It’s a race between the growing power of the technology and the growing wisdom we need to manage it.”

‘Future is Ours to Shape’

There is a second Swede in this story, and even more than Bostrom, he’s the person driving the conversation. His name is Max Tegmark. He’s a charismatic 48-year-old professor in the physics department at the Massachusetts Institute of Technology. He’s also a founder of something called the Future of Life Institute, which has been doling out Elon Musk’s money for research on making A.I. safer.

Tegmark is something of a physics radical, the kind of scientist who thinks there may be other universes in which not only the speed of light and gravity are different but the mathematical underpinnings of reality are different. Tegmark and Bostrom are intellectual allies. In Tegmark’s recent book, “Our Mathematical Universe: My Quest for the Ultimate Nature of Reality,” he writes about meeting Bostrom at a conference in 2005 in California:

“After some good wine, our conversation turned to doomsday scenarios. Could the Large Hadron Collider create a miniature black hole that would end up gobbling up Earth? Could it create a ‘strangelet’ that could catalyze the conversion of Earth into strange quark matter?”

In addition to taking the what-could-go-wrong questions seriously, Tegmark and Bostrom entertain optimistic scenarios. Perhaps, they say, Earth is the only planet in the universe that harbors intelligent life. We have a chance to take this startling phenomenon of intelligence and spread it to the stars — if we don’t destroy ourselves first with runaway technology.

“The future is ours to shape. I feel we are in a race that we need to win. It’s a race between the growing power of the technology and the growing wisdom we need to manage it. Right now, almost all the resources tend to go into growing the power of the tech,” Tegmark said.

The MIT cosmologist Max Tegmark is among those who believe that artificial intelligence could be either the best or worst thing to happen to the human race. He and fellow scientists and thinkers have founded the Future of Life Institute, which is pouring millions of dollars into research on A.I. safety. Tegmark has also opposed development of autonomous weaponized machines.

In April 2014, 33 people gathered in Tegmark’s home to discuss existential threats from technology. They decided to form the Future of Life Institute. It would have no paid staff members. Tegmark persuaded numerous luminaries in worlds of science, technology and entertainment to add their names to the cause. Skype founder Jaan Tallinn signed on as a co-founder. Actors Morgan Freeman and Alan Alda joined the governing board.

Tegmark put together an op-ed about the potential dangers of machine intelligence, lining up three illustrious co- authors: Nobel laureate physicist Frank Wilczek, artificial intelligence researcher Stuart Russell, and the biggest name in science, Stephen Hawking. Hawking’s fame is like the midday sun washing out every other star in the sky, and Tegmark knew that the op-ed would be viewed as an oracular pronouncement from the physicist

The piece, which ran in the Huffington Post and in the Independent in Britain, was a brief,

breezy tract that included a tutorial on the idea of the Singularity and a dismayed conclusion that experts weren't taking the threat of runaway A.I. seriously. A.I., the authors wrote, is "potentially the best or worst thing ever to happen to humanity."

"Stephen Hawking Says A.I. Could Be Our 'Worst Mistake In History,' " one online science news site reported. And CNBC declared: "Artificial intelligence could end mankind: Hawking." So that got everyone's attention.

Tegmark's next move was to organize an off-the-record conference of big thinkers to discuss A.I. While the Boston area went into a deep freeze in January of this year, about 70 scientists and academics, led by Tegmark, convened in Puerto Rico to discuss the existential threat of machine intelligence. Their model was the historic conference on recombinant DNA research held in Asilomar, Calif., in 1975, which resulted in new safeguards for gene splicing.

Musk, the founder of Tesla and SpaceX, joined the group in Puerto Rico. On the final night of the conference, he pledged $10 million for research on lowering the threat from A.I.

"With artificial intelligence, we are summoning the demon," Musk had said earlier, a line that sent Twitter into a tizzy.

In the months that followed, 300 teams of researchers sent proposals for ways to lower the A.I. threat. Tegmark says the institute has awarded 37 grants worth $7 million.

"What we're doing every day today is producing super stupid entities that make mistakes."

Humans vs. Machines

Reality check. More than half a century of research on artificial intelligence has yet to produce anything resembling a conscious, willful machine. We still control this technology. We can unplug it.

Just down Vassar Street from Tegmark's office is MIT's Computer Science and Artificial Intelligence Laboratory, where robots are aplenty. Daniela Rus, the director, is an inventor who just nabbed $25 million in funding from Toyota to develop a car that will never be involved in a collision.

Is she worried about the Singularity?"It rarely comes up," Rus said. "It's just not something I think about."

With a few exceptions, most full-time A.I. researchers think the Bostrom-Tegmark fears are premature. A widely repeated observation is that this is like worrying about overpopulation on Mars. Rus points out that robots are better than humans at crunching numbers and lifting heavy loads, but humans are still better at fine, agile motions, not to mention creative, abstract thinking.

"The progress has not been as steady as people say, and the machine skills are really far from being ready to match our skills," she said. "There are tasks that are very easy for humans — clearing your dinner table, loading the dishwasher, cleaning up your house — that are surprisingly difficult for machines."

Rus makes a point about self-driving cars: They can't drive just anywhere. They need precise maps and relatively predictable situations. She believes, for example, that they couldn't handle Washington's Dupont Circle.

In Dupont Circle, vehicles and pedestrians muddle their way forward through a variety of interpersonal signals that a machine could not interpret, she said. Self-driving cars struggle with heavy traffic, she said, and even rain and snow are a problem. So imagine trying to understand hand gestures from road crews and other drivers.

"There's too much going on," Rus said. "We don't have the right sensors and algorithms to characterize very quickly what happens in a congested area, and to compute how to react."

Too Much Power

The future is implacably murky when it comes to technology; the smartest people on the planet fail to see what's coming. For example, many of the great sages of the modern era didn't anticipate that computers would get smaller rather than bigger.

Anyone looking for something to worry about in the near future might want to consider the opposite of superintelligence: superstupidity.

In our increasingly technological society, we rely on complex systems that are vulnerable to failure in complex and unpredictable ways. Deepwater oil wells can blow out and take months to be resealed. Nuclear power reactors can melt down. Rockets can explode. How might intelligent machines fail — and how catastrophic might those failures be?

Often there is no one person who understands exactly how these systems work or are operating at any given moment. Throw in elements of autonomy, and things can go wrong quickly and disastrously.

Superintelligence hasn't yet arrived, but clever, interactive robots have. Roboticist Cynthia Breazeal of MIT is the human mind behind a social robot, named Jibo, that is embedded with artificial intelligence software and sensors to enable it to recognize, and interact with, multiple people at once. Breazeal said Jibo will become commercially available in early 2016. Click here to read more about Jibo and other social robots.

Such was the case with the "flash crash" in the stock market in 2010, when, in part because of automated, ultra- fast trading programs, the Dow Jones industrial average dropped almost 1,000 points within minutes before rebounding. "What we're doing every day today is producing super stupid entities that make mistakes," argues Boris Katz, another artificial intelligence researcher at MIT.

"Machines are dangerous because we are giving them too much power, and we give them power to act in response to sensory input. But these rules are not fully thought through, and then sometimes the machine will act in the wrong way," he said."But not because it wants to kill you."

A living legend of the A.I. field and the MIT faculty is Marvin Minsky, 88, who helped found the field in the 1950s. It was his generation that put us on this road to the age of smart machines.

Minsky, granting an interview in the living room of his home a few miles from campus, flashed an impish smile when asked about the dangers of intelligent machines. “I suppose you could write a book about how they’ll save us,” he said. “It just depends upon what dangers appear.”

Taking on Autonomous Weapons

The A.I. debate is likely to remain tangled in uncertainties and speculation. In theory, as that original Huffington Post op-ed stated, there’s no theoretical limit to machine intelligence — “no physical law precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains.”

But the academic and scientific establishment is not convinced that A.I. is an imminent threat.

Tegmark and his Future of Life allies decided this summer to take on a related but more urgent issue: the threat of autonomous weaponized machines.

Tegmark teamed with Stuart Russell, the artificial intelligence researcher, on an open letter calling for a ban on such weapons. Once again, they got Hawking to sign it, along with Musk, Bostrom, and about 14,000 other scientists and engineers. On July 28, they formally presented the letter at an A.I. conference in Buenos Aires.

Russell said it took him five minutes of Internet searching to figure out how a very small robot — a microbot — could use a shaped charge to “blow holes in people’s heads.” A microrifle, he said, could be used to “shoot their eyes out.”

“You’d have large delivery ships that would dump millions of small flying vehicles, probably even insect-sized, the smallest you could get away with, and still kill a human being,” Russell said.

“I actually think it would be a huge tragedy if machine superintelligence were never developed. That would be a failure mode for our Earth-originating intelligent civilization.”

Unlocking Possibilities

After Nick Bostrom’s lecture in Santa Fe, held at the New Mexico School for the Deaf, he went to a book-signing event across town at the School for Advanced Research. His book is a meticulously reasoned, rather dense tract titled, “Superintelligence: Paths, Dangers, Strategies.”

Bostrom, standing at the edge of a courtyard, held forth amid a small cluster of party guests. Then he sat down for an hour-long interview. Reserved, intensely focused on his ideas, the 42-year-old Bostrom seemed apprehensive about whether his ideas could be fully grasped by someone who is not an academic philosopher. He was distracted by the possibility that a gnat, or fly, or some such insect had invaded his water glass.

Asked if there was something he now wishes he had done differently with his book, he said he should have been clear that he supports the creation of superintelligence. Unsurprisingly, most readers missed that key point.

“I actually think it would be a huge tragedy if machine superintelligence were never developed,” he said. “That would be a failure mode for our Earth-originating intelligent civilization.”

In his view, we have a chance to go galactic — or even intergalactic — with our intelligence. Bostrom, like Tegmark, is keenly aware that human intelligence occupies a minuscule space in the grand scheme of things. The Earth is a small rock orbiting an ordinary star on one of the spiral arms of a galaxy with hundreds of billions of stars. And at least tens of billions of galaxies twirl across the known universe.

Artificial intelligence, Bostrom said, “is the technology that unlocks this much larger space of possibilities, of capabilities, that enables unlimited space colonization, that enables uploading of human minds into computers, that enables intergalactic civilizations with planetary-size minds living for billions of years.”

There’s a bizarre wrinkle in Bostrom’s thinking. He thinks a superior civilization would possess essentially infinite computing power. These superintelligent machines could do almost anything, including create simulated universes that include programs that precisely mimic human consciousness, replete with memories of a person’s history — even though all this would be entirely manufactured by software, with no real-world, physical manifestation.

Bostrom goes so far as to say that unless we rule out the possibility that a machine could create a simulation of a human existence, we should assume that it is overwhelmingly likely that we are living in such a simulation.

“I’m not sure that I’m not already in a machine,” he said calmly.