Anxious about A.I.

Some of the world's top scientists and philosophers believe that intelligent machines pose a threat to humanity. Journalist Joel Achenbach asked them to explain why they are so worried.

THE WORLD'S SPOOKIEST philosopher is Nick Bostrom, a thin, soft-spoken Swede. Of all the people worried about runaway artificial intelligence, 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 earlier this year. He added, maintaining his tone of understatement, "I think that would be a low-value future."

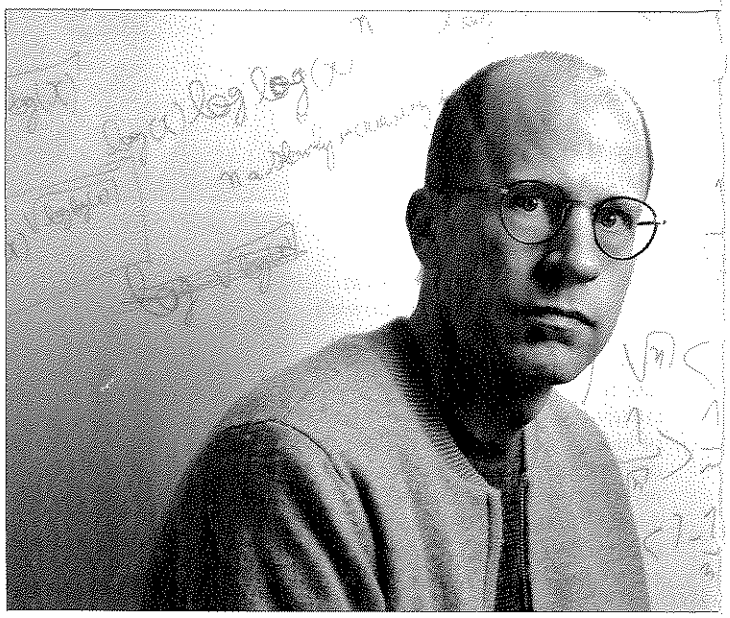
Bostrom's underlying concerns about machine intelligence, unintended consequences, and potentially malevolent computers have gone mainstream. You can't attend a technology conference these days without someone bringing up A.I. anxiety. It hovers over the tech conversation with the high-pitched whine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

People will tell you that even Stephen Hawking is worried about it. And Bill Gates. And that Tesla founder 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.



Philosopher Nick Bostrom believes A.I. must be treated with vigilance.

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 will actually happen is the subject of robust debate. Bostrom supports the research but worries that sufficient safeguards are not in place. Imagine, he 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 being born.

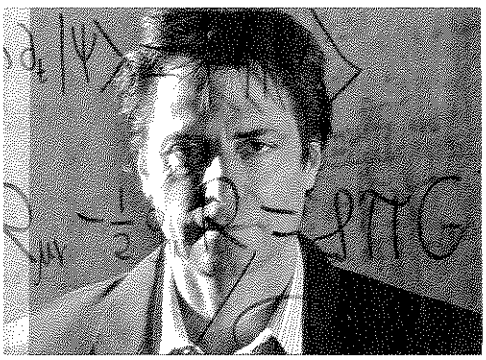
Or imagine, Bostrom says, that superintelligent machines are programmed to ensure that whatever they do will make humans smile. They may then decide that they should implant electrodes into the facial muscles of all people to keep us smiling.

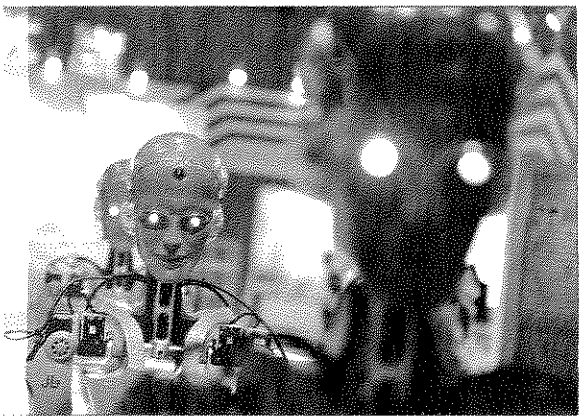
Bostrom isn't saying this will happen. These are thought experiments. His big-picture idea is that, just in the past couple of centuries, we've seen astonishing changes in the human population and economic prosperity. In Bostrom's view, our modern existence is an anomaly—one created largely by technology.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

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

Tom Pilston/The Washington Post





Physicist Max Tegmark heads the Future of Life Institute.

may be other universes in which not only the speed of light and gravity are different but also the mathematical underpinnings of reality. Tegmark and Bostrom are intellectual allies.

"The future is ours to shape," Tegmark said. "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."

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 the 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. The piece was a brief, breezy tract that included 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."

CNBC declared: "Artificial intelligence could end mankind: Hawking."

That got everyone's attention.

JUST DOWN THE street from Tegmark's office is MIT's Computer Science and Artificial Intelligence Lab, where robots are aplenty. Director Daniela Rus is an inventor who just nabbed $25 million 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.

She makes a point about self-driving cars: They can't drive just anywhere. They need precise maps and predictable situations. Self-driving cars struggle with heavy traffic, she said, and even rain and snow are a problem. Imagine them trying to understand hand gestures from other drivers. "There's too much going on," Rus said.

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. 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.

Such was the case with the "flash crash" in the stock market in 2010, when, in part because of automated, ultrafast 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 superstupid 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."

"I ACTUALLY THINK it would be a huge tragedy if machine superintelligence were never developed," Bostrum 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. 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 believes a superior civilization would possess essentially infinite computing power. These superintelligent machines could do almost anything, even create simulated universes that included programs that precisely mimicked 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.

Excerpted from an article that originally appeared in The Washington Post. *Reprinted with permission.*