Nick Bostrom, a philosopher who directs the Future of Humanity Institute at the University of Oxford, describes the following scenario in his book Superintelligence, which has prompted a great deal of debate about the future of artificial intelligence. Imagine a machine that we might call a "paper-clip maximizer"—that is, a machine programmed to make as many paper clips as possible. Now imagine that this machine somehow became incredibly intelligent. Given its goals, it might then decide to create new, more efficient paper-clip-manufacturing machines—until, King Midas style, it had converted essentially everything to paper clips.

No worries, you might say: you could just program it to make exactly a million paper clips and halt. But what if it makes the paper clips and then decides to check its work? Has it counted correctly? It needs to become smarter to be sure. The superintelligent machine manufactures some as-yet-uninvented raw-computing material (call it "computronium") and uses that to check each doubt. But each new doubt yields further digital doubts, and so on, until the entire earth is converted to computronium. Except for the million paper clips.

Things Reviewed

Bostrom does not believe that the paper-clip maximizer will come to be, exactly; it's a thought experiment, one designed to show how even careful system design can fail to restrain extreme machine intelligence. But he does believe that superintelligence could emerge, and while it could be great, he thinks it could also decide it doesn't need humans around. Or do any number of other things that destroy the world. The title of chapter 8 is: "Is the default outcome doom?"

If this sounds absurd to you, you're not alone. Critics such as the robotics pioneer Rodney Brooks say that people who fear a runaway AI misunderstand what computers are doing when we say they're thinking or getting smart. From this perspective, the putative superintelligence Bostrom describes is far in the future and perhaps impossible.

Yet a lot of smart, thoughtful people agree with Bostrom and are worried now. Why?

Volition

The question "Can a machine think?" has shadowed computer science from its beginnings. Alan Turing proposed in 1950 that a machine could be taught like a child; John McCarthy, inventor of the programming language LISP, coined the term "artificial intelligence" in 1955. As AI researchers in the 1960s and 1970s began to use computers to recognize images, translate between languages, and understand instructions in normal language and not just code, the idea that computers would eventually develop the ability to speak and think—and thus to do evil—bubbled into mainstream culture. Even beyond the oft-referenced HAL from *2001: A Space Odyssey*, the 1970 movie *Colossus: The Forbin Project* featured a large blinking mainframe computer that brings the world to the brink of nuclear destruction; a similar theme was explored 13 years later in *WarGames*. The androids of 1973's *Westworld* went crazy and started killing.

When AI research fell far short of its lofty goals, funding dr ied up to a trickle, beginning long "AI winters." Even so, the torch of the intelligent machine was carried forth in the 1980s and '90s by sci-fi authors like Vernor Vinge, who popularized the concept of the singularity; researchers like the roboticist Hans Moravec, an expert in computer vision; and the engineer/entrepreneur Ray Kurzweil, author of the 1999 book *The Age of Spiritual Machines*. Whereas Turing had posited a humanlike intelligence, Vinge, Moravec, and Kurzweil were thinking bigger: when a computer became capable of independently devising ways to achieve goals, it would very likely be capable of introspection—and thus able to modify its software and make itself more intelligent. In short order, such a computer would be able to design its own hardware.

As Kurzweil described it, this would begin a beautiful new era. Such machines would have the insight and patience (measured in picoseconds) to solve the outstanding problems of nanotechnology and spaceflight; they would improve the human condition and let us upload our consciousness into an immortal digital form. Intelligence would spread throughout the cosmos.

You can also find the exact opposite of such sunny optimism. Stephen Hawking has warned that because people would be unable to compete with an advanced AI, it "could spell the end of the human race." Upon reading *Superintelligence*, the entrepreneur Elon Musk tweeted: "Hope we're not just the biological boot loader for digital superintelligence. Unfortunately, that is increasingly probable." Musk then followed with a $10 million grant to the Future of Life Institute. Not to be confused with Bostrom's center, this is an organization that says it is "working to mitigate existential risks facing humanity," the ones that could arise "from the development of human-level artificial intelligence."

No one is suggesting that anything like superintelligence exists now. In fact, we still have nothing approaching a general-purpose artificial intelligence or even a clear path to how it could be achieved. Recent advances in AI, from automated assistants such as Apple's Siri to Google's driverless cars, also reveal the technology's severe limitations; both can be thrown off by situations that they haven't encountered before. Artificial neural networks can learn for themselves to recognize cats in photos. But they must be shown hundreds of thousands of examples and still end up much less accurate at spotting cats than a child.

This is where skeptics such as Brooks, a founder of iRobot and Rethink Robotics, come in. Even if it's impressive—relative to what earlier computers could manage—for a computer to recognize a picture of a cat, the machine has no volition, no sense of what cat-ness is or what else is happening in the picture, and none of the countless other insights that humans have. In this view, AI could possibly lead to intelligent machines, but it would take much more work than people like Bostrom imagine. And even if it could happen, intelligence will not necessarily lead to sentience. Extrapolating from the state of AI today to suggest that superintelligence is looming is "comparable to seeing more efficient internal combustion engines appearing and jumping to the conclusion that warp drives are just around the corner," Brooks wrote recently on Edge.org. "Malevolent AI" is nothing to worry about, he says, for a few hundred years at least.

"Superintelligence: **Paths, Dangers, Strategies**"
------Nick Bostrom

Insurance Policy

Even if the odds of a superintelligence arising are very long, perhaps it's irresponsible to take the chance. One person who shares Bostrom's concerns is Stuart J. Russell, a professor of computer science at the University of California, Berkeley. Russell is the author, with Peter Norvig (a peer of Kurzweil's at Google), of *Artificial Intelligence: A Modern Approach*, which has been the standard AI textbook for two decades.

"There are a lot of supposedly smart public intellectuals who just haven't a clue," Russell told me. He pointed out that AI has advanced tremendously in the last decade, and that while the public might understand progress in terms of Moore's Law (faster computers are doing more), in fact recent AI work has been fundamental, with techniques like deep learning laying the groundwork for computers that can automatically increase their understanding of the world around them. Because Google, Facebook, and other companies are actively looking to create an intelligent, "learning" machine, he reasons, "I would say that one of the things we ought not to do is to press full steam ahead on building superintelligence without giving thought to the potential risks. It just seems a bit daft." Russell made an analogy: "It's like fusion research. If you ask a fusion researcher what they do, they say they work on containment. If you want unlimited energy you'd better contain the fusion reaction." Similarly, he says, if you want unlimited intelligence, you'd better figure out how to align computers with human needs. ■

*How did people think about **Artificial Intellgence***

1950
"Machine could be thought like a **CHILD**"
------Alan Turing

1955
"computers could **UNDERSTAND** human intelligence"
------John McCarthy

1960s & 1970s
Scientists believed computers would develop the ability to **SPEAK AND THINK**.

1980s & 1990s
"Computers would be able to **INTROSPECT**"
------Hans Moravec

2014
"AI could spell the **END** of the human ace"
------Stephen Hawking

2015
An open letter:
"avoiding p otential **PITFALLS** of AI"
------over 8,000 AI researchers