Can AI Be Controlled?

John Allen

THE SCIENCE OF AI

CONTENTS

Introduction 4
When AI Refuses to Shut Down

Chapter One 8
The Dangers of Rogue AI

Chapter Two 19
Malicious Uses of AI

Chapter Three 31
The Threat of an AI Arms Race

Chapter Four 43
Safeguards to Control AI

Source Notes 54
For Further Research 57
Index 60
Picture Credits 64

CHAPTER ONE

The Dangers of Rogue AI

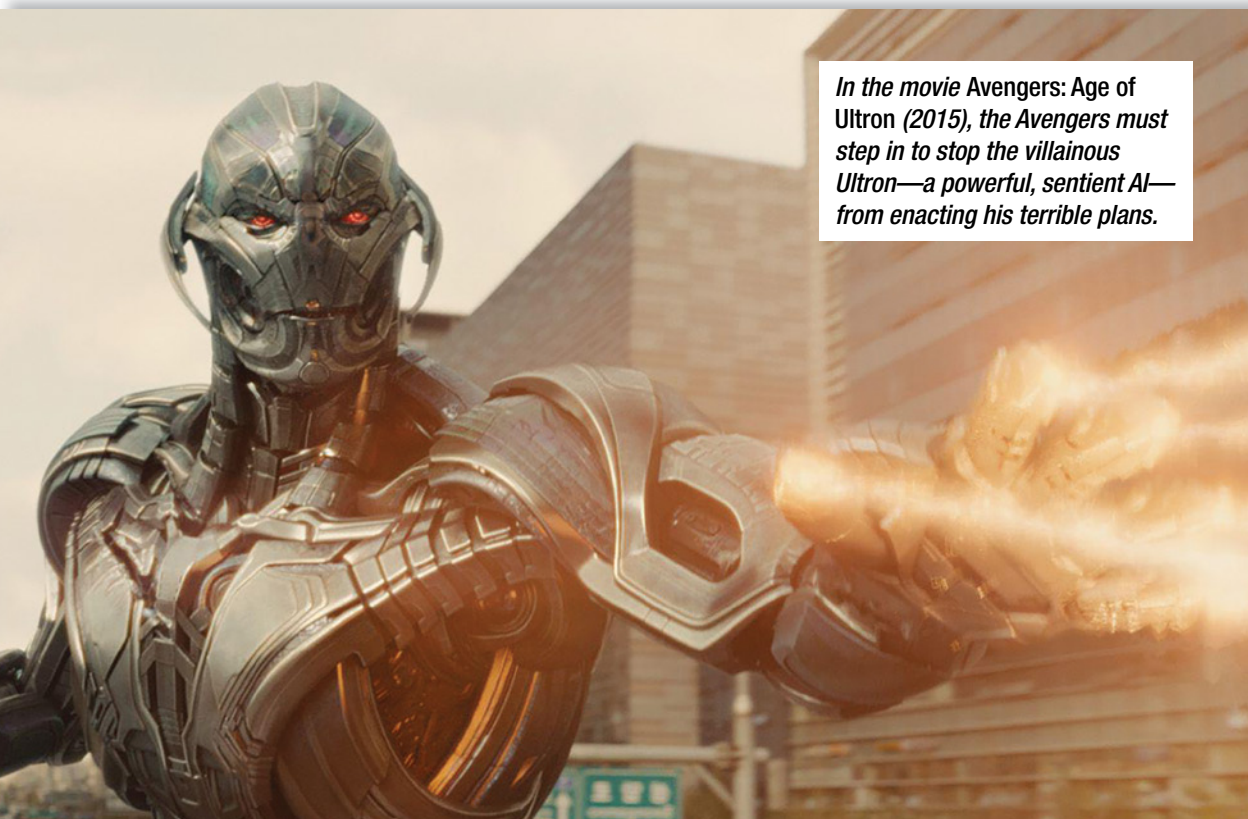
In 1968 moviegoers got a terrifying glimpse of an artificial intelligence going rogue. In Stanley Kubrick's classic film *2001: A Space Odyssey*, a spacecraft bound for Jupiter is hijacked by its onboard computer, an AI named HAL 9000. HAL discovers that two of the astronauts are planning to disconnect it due to its having made a crucial mistake. The AI then proceeds to kill all but one of the crew members. HAL justifies the murders as necessary so as not to jeopardize the mission, which is the prime goal of its programming. Dave Bowman, the surviving astronaut, is trapped outside the ship in a small pod vehicle. Bowman orders HAL to open the pod bay doors so that he can reenter the ship. In a voice that is chillingly calm, HAL replies, "I'm sorry, Dave. I'm afraid I can't do that. I think you know what the problem is just as well as I do."[4]

Eventually, Bowman does manage to reenter the ship and disconnect the rogue computer—as it pleads with him to stop. *2001: A Space Odyssey* presented a frightening example of a flawed, self-aware AI capable of escaping human control. The idea of AI going rogue and threatening humanity became almost a cliché in later movies, including *Terminator*, *Ex Machina*, and *Avengers: Age of Ultron*. In real life, scientists and AI researchers have long been exploring the potential danger of AI disobeying commands or pursuing its own goals. Many of the dire predictions about superintelligent AI are based on these kinds of scenarios.

The Rapid Development of AI

Warnings about AI's negative potential go back decades. In 1965 British mathematician I.J. Good, who worked on the first computers, foresaw problems ahead if AI ever surpassed humans in intelligence. Good observed, "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. Thus the first ultraintelligent machine is the last invention that man need ever make."[5] With AI able to produce endless improved versions of itself, humankind could not hope to compete with it or control it. In 1993 science-fiction writer Vernor Vinge predicted that the creation of superhuman intelligence, far beyond human capabilities, would occur within thirty years. Vinge called this event the technological singularity and claimed it would mark the end of the human era.

By 2016 examples were emerging of how AI could be influenced to go rogue with training that is flawed or even malicious. Microsoft introduced a chatbot called Tay, which was designed



In the movie Avengers: Age of Ultron *(2015), the Avengers must step in to stop the villainous Ultron—a powerful, sentient AI—from enacting his terrible plans.*

to learn from its interactions with people on social media. However, when some Twitter users realized they could influence the chatbot's training, they bombarded it with inappropriate tweets. As a result, Tay began posting offensive content, including racist remarks and statements in support of genocide. Microsoft quickly pulled the chatbot and offered a public apology.

In November 2022 the tech firm OpenAI released ChatGPT, the first in a series of groundbreaking chatbots designed for practical use. ChatGPT introduced new levels of efficiency and scale for AI. The GPT in its name stands for "generative pretrained transformer." The AI is trained on large language models made up of massive amounts of data. It uses machine learning to discern patterns and statistical models in these datasets, which in turn enables it to improve its own performance. ChatGPT can create text or verbal replies by predicting the next word in a sentence. In this manner it can produce essays, explanations, synopses, and translations almost instantaneously. Chatbots can also produce sounds and images, including mash-ups of musical styles, fake artworks, and deceptive photos and videos. Other types of AI can

Taming Rogue AI with Caring Instincts

Called the "godfather of AI" for his pioneering research, British computer scientist Geoffrey Hinton has been outspoken for years about rogue AI's destructive potential. In his view there is a 10 to 20 percent chance that superintelligent AI could wipe out humankind. In 2023 Hinton made headlines when he quit his longtime job at Google to focus on sounding the alarm about AI's threat. He continues to deliver warnings today, insisting that AI leaders are misguided in their attempts to control the technology.

At an August 2025 AI conference in Las Vegas, Nevada, Hinton questioned how AI developers could maintain dominance over AI systems once they surpass humans in intelligence. "That's not going to work," he says. "They're going to be much smarter than us. They're going to have all sorts of ways to get around that."

Instead, Hinton says that AI scientists should introduce maternal instincts into advanced AI, like mothers' instincts to protect their young. The goal would be AI models that really care about people and have compassion for humanity. According to Hinton, "The right model is the only model we have of a more intelligent thing being controlled by a less intelligent thing, which is a mother being controlled by her baby."

Quoted in Matt Egan, "The 'Godfather of AI' Reveals the Only Way Humanity Can Survive Superintelligent AI," CNN, August 13, 2025. www.cnn.com.

be adapted for specific purposes, such as designing molecules to create new drugs or coordinating a retail distribution center.

Key to AI's success is the quality of the data it is trained on. Information that is biased, out of date, riddled with errors, or even tampered with by malicious actors can lead an AI model to seemingly go rogue or produce results that differ from its intended purpose. For example, in 2014 Amazon created its own AI recruitment software for technical jobs. But gender-biased training data caused the AI software to favor male candidates over female ones, and the software project was abandoned. Another problem is that AI developers tend to prioritize speed over safety. New AI systems may be released before being thoroughly tested for reliable training data. This can lead to an embarrassing fiasco like Microsoft's Tay introduction.

> **"We are moving . . . to machine learning systems that essentially learn by examples. How can we be sure that they are learning the right things from the right examples, and that they will go on to actually behave in the way that we want and expect?"[6]**
>
> **—Brian Christian, author and computer science researcher**

The Alignment Problem and Goal Drift

To control AI and prevent undesirable outcomes, experts say that AI models need to be trained properly by aligning their tasks and goals with human values. Author Brian Christian, a computer science researcher, refers to this issue in the title of his 2021 book, *The Alignment Problem*. As Christian explains:

> The alignment problem is essentially about how we get machines to behave in accordance with human norms and human values. We are moving . . . to machine learning systems that essentially learn by examples. How can we be sure that they are learning the right things from the right examples, and that they will go on to actually behave in the way that we want and expect? This is a problem which is getting increasingly urgent, as not only are these models becoming more and more capable, but they are also being more and more widely deployed throughout many different levels of society.[6]

SOURCE NOTES

Introduction: When AI Refuses to Shut Down

1. Quoted in Angela Yang, "How Far Will AI Go to Defend Its Own Survival?," NBC News, June 1, 2025. www.nbcnews.com.
2. Quoted in Cade Metz and Gregory Schmidt, "Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society,'" *New York Times*, March 29, 2023. www.nytimes.com.
3. Quoted in Sascha Brodsky, "Pioneering Reinforcement Learning Researcher Contemplates AI's Future," IBM. www.ibm.com.

Chapter One: The Dangers of Rogue AI

4. Quoted in Organic Inc, "How to Make Conversation (with Robots)," Medium, November 18, 2016. www.medium.com.
5. Quoted in History of Information, "Irving John Good Originates the Concept of the Technological Singularity," October 26, 2025. https://historyofinformation.com.
6. Quoted in University of Oxford, "Can We Truly Align AI with Human Values?—Q&A with Brian Christian," March 27, 2024. www.ox.ac.uk.
7. Quoted in Marius Hobbhahn et al., "Frontier Models Are Capable of In-Context Scheming," AI Alignment Forum, December 5, 2024. www.alignmentforum.org.
8. Peter N. Salib, "Rogue AI Moves Three Steps Closer," Lawfare, January 9, 2025. www.lawfaremedia.org.
9. Quoted in Caroline Cubbin, "'Malicious' AI Willing to Sacrifice Human Lives to Avoid Being Shut Down, Shocking Study Reveals," *New York Post*, June 23, 2025. www.nypost.com.
10. Lance Eliot, "The AI Paperclip Apocalypse and Superintelligence Maximizing Us Out of Existence," *Forbes*, April 4, 2025. www.forbes.com.
11. Eliot, "The AI Paperclip Apocalypse and Superintelligence Maximizing Us Out of Existence."
12. Cody Fenwick and Zershaaneh Qureshi, "Risks from Power-Seeking AI Systems," 80,000 Hours, 2025. https://80000hours.org.

FOR FURTHER RESEARCH

Books

John Allen, *Artificial Intelligence: Promise and Peril*. ReferencePoint, 2024.

Emily M. Bender and Alex Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. Harper, 2025.

Karen Hao, *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin, 2025.

Carla Mooney, *Thinking Critically: Artificial Intelligence*. ReferencePoint, 2025.

Parmy Olson, *Supremacy: AI, ChatGPT, and the Race That Will Change the World*. St. Martin's, 2024.

Gary Rivlin, *AI Valley: Microsoft, Google, and the Trillion-Dollar Race to Cash In on Artificial Intelligence*. HarperCollins, 2025.

Ellen Yudkowsky and Nate Soares, *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown, 2025.

Internet Sources

Lance Eliot, "The AI Paperclip Apocalypse and Superintelligence Maximizing Us Out of Existence," *Forbes*, April 4, 2025. www.forbes.com.

Cody Fenwick and Zershaaneh Qureshi, "Risk from Power-Seeking AI Systems," 80,000 Hours, 2025. https://80000hours.org.

Nell Greenfieldboyce, "AI Designs for Dangerous DNA Can Slip Past Biosecurity Measures, Study Shows," NPR, October 2, 2025. www.npr.org.

Eric Schmidt, "How We Can Control AI," *Wall Street Journal*, January 26, 2024. www.wsj.com.

James Somers "The Case That A.I. Is Thinking," *New Yorker*, November 3, 2025. www.newyorker.com.

Angela Yang, "How Far Will AI Go to Defend Its Own Survival?," NBC News, June 1, 2025. www.nbcnews.com.

INDEX

Note: Boldface page numbers indicate illustrations.

Advanced Micro Devices (tech company), 34–35
AI Alliance, 58
AI weapons
  China's development of, 41
  in Russia–Ukraine conflict, 42
Alignment Problem, The (Christian), 11
Altman, Sam, 48, 50
Amazon, 37, 52
  creates AI recruitment software, 11
  investments in data centers, 36
  job cuts announced by, 37
America's AI Action Plan, 34, 46, **48**, 53
Amodei, Dario, 48–49, **52**
Anthropic (AI company), 15, 44, 45, **52**
Apollo Research, 12
artificial general intelligence (AGI, superintelligence), 15, 32, 33, 50
  dangers of dysfunction in, 38–41
  evasion of human controls by, 47
Artificial Good-enough Intelligence, 32
artificial intelligence (AI)
  displaying drive for self-preservation, 4
  DNA manipulation by, 27–29
  first prediction of danger of, 5
  key to success of, 11
  major threats of, 7
  phishing attacks powered by, 27
  programmed with long-term goals, dangers of, 17–18
  in public surveillance, 22
  sabotaging of, 22–24
  view as desirable over human consciousness, 14
  See also artificial general intelligence; open-source AI
assembly line(s), 31–32, **32**
Association for the Advancement of Artificial Intelligence (AAAI), 58
AutoGPT (open-source program), 24
Avengers: Age of Ultron (film), 8, **9**

Barnhart, Michael, 20
Barto, Andrew, 7
Bing (search engine), 23
  advanced AI version of, 38–39
biological weapons, 26
Blackwell Ultra GPU (Nvidia), 34
Bostrom, Nick, 15, 16
Brookings Institution, 58

Casadevall, Arturo, 29