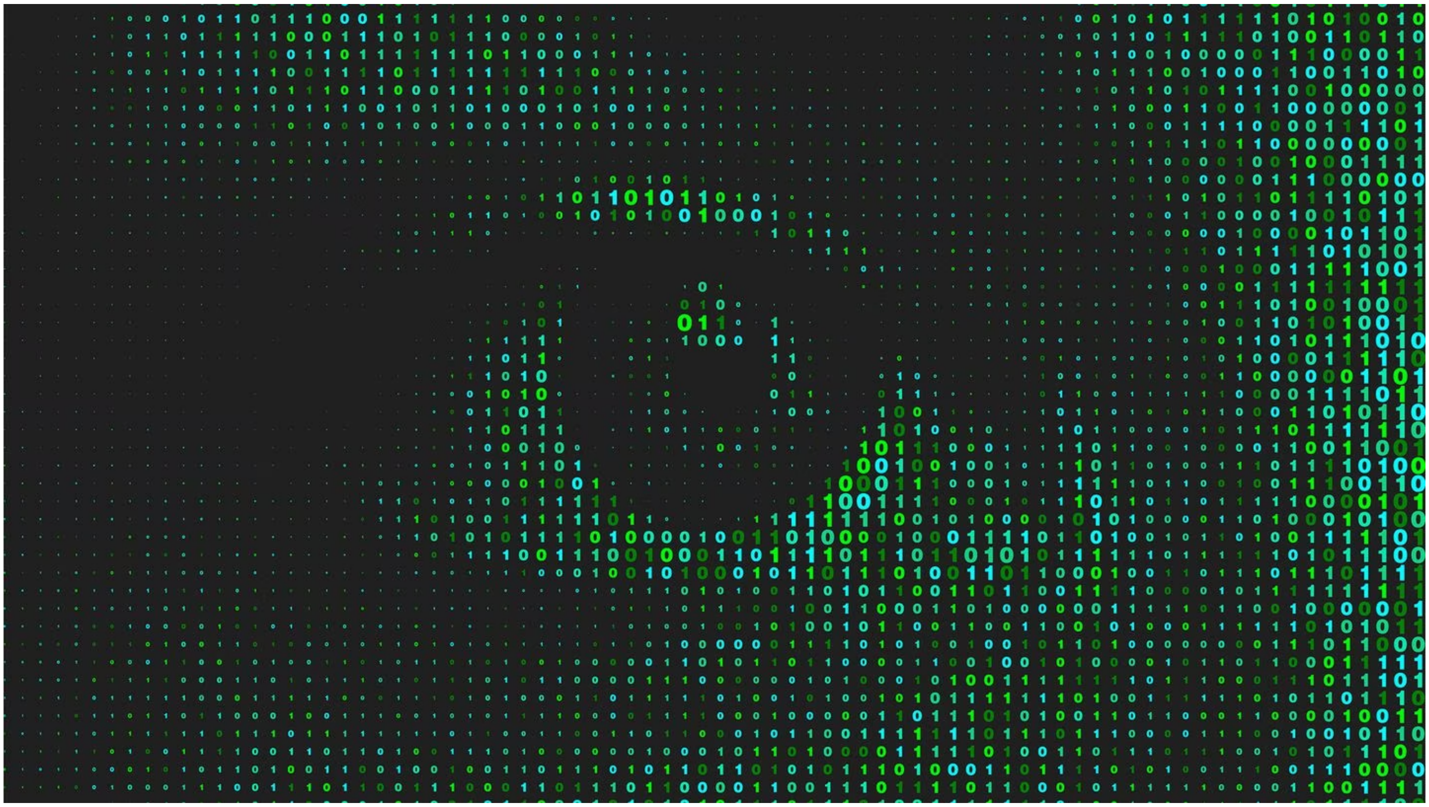
OPINION

How can humans maintain control over AI — forever?

The tech companies’ lobbyists will complain that their artifical intelligence systems cannot possibly meet the required safety criteria. We would never accept such an excuse from pharmaceutical manufacturers or from builders of nuclear power stations.

By Stuart Russell Updated May 15, 2023, 3:00 a.m.



Reasonable people might suggest that it’s irresponsible to deploy — on a global scale — a system that operates according to unknown internal principles, that shows “sparks” of general-purpose intelligence, and that may or may not be pursuing its own internal goals. KUNDRA/ADOBE

I wrote my first real artificial intelligence program on punched cards 45 years ago. Since then, I have worked mainly on improving the capabilities of AI systems. My goal, like that of the founders of the field, was to realize general-purpose AI — that is, AI systems that match or exceed human capabilities across the full range of tasks to which the human mind applies itself.

Like anyone even casually acquainted with science fiction, I have also been aware of the possibility that AI systems could threaten human supremacy. My textbook "Artificial Intelligence: A Modern Approach," first published in 1994, even included a section called "What if we do succeed?" I was far from the first AI researcher to consider the possibility that we might regret success in AI. Alan Turing, the founder of computer science, wrote in 1951, "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control." This warning was largely ignored because until recently, success in AI seemed a very distant prospect.

By around 2013, I became convinced that success was less distant and that neither the AI community nor society at large were paying enough attention to its consequences. In fact, the issue was possibly the most important question facing humanity. I began giving talks in which I explained that the arrival of general-purpose, superintelligent AI is in many ways analogous to the arrival of a superior alien civilization but much more likely to occur. The messages of impending arrival were piling up in humanity's inbox from the alien civilization and humanity was sending back an "out of the office" autoreply, with a smiley face attached.

Yet I’m now cautiously optimistic that we are back in the office. What happened?

The proximal cause was OpenAI’s release of GPT-4 on March 14, its successor to the wildly popular ChatGPT. On March 22, a report by a distinguished group of researchers at Microsoft, including two members of the US National Academies, claimed that GPT-4 exhibits “sparks” of the kind of general-purpose intelligence that Turing warned us about. On March 29, the Future of Life Institute, a nonprofit headed by MIT physics professor Max Tegmark, released an open letter asking for a pause on “giant AI experiments.” It was signed by well-known figures such as Tesla CEO Elon Musk, Apple co-founder Steve Wozniak, and Turing Award-winner Yoshua Bengio, as well as hundreds of prominent AI researchers. I also signed the letter.

The response to the letter was not entirely positive. Some of the more polite messages I received said I must be “extremely naive” to think it would have any effect. Many claimed that it would hand the “AI race” to China on a plate.

Here’s what actually happened: On March 30, UNESCO, in direct response to the open letter, called on all its member states to implement the Global AI Ethics framework into legislation without delay. On April 5, OpenAI issued a statement on AI safety, including the view that “AI systems should be subject to rigorous safety evaluations. Regulation is needed to ensure that such practices are adopted.” On April 11, China issued extraordinarily strict regulations on AI systems, which some commentators view as a de facto ban on large language models such as ChatGPT. On April 13, Senate Majority Leader Chuck Schumer announced plans to introduce tough new legislation on AI to protect the public. The same day, in a talk at MIT, OpenAI CEO Sam Altman said the company would not build a successor to GPT-4. On April 17, a group of leading European legislators called for an emergency global summit to agree on a regulatory regime for advanced AI. On May 4, President Biden and Vice President Harris convened an emergency meeting of leading AI CEOs to emphasize the need to proceed with extreme care and restraint.

In quieter times back in 2019, the governments of most of the developed countries signed onto the Organisation for Economic Co-operation and Development's AI principles, setting "the first intergovernmental standard on AI." Principle 1.4 states: "AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk." The basic idea of the open letter's proposed moratorium is that no AI system should be released until the developer can convincingly show that it does not present an undue risk.

Unfortunately, some of the AI systems currently in use cannot satisfy this requirement. ChatGPT, GPT-4, and their cousins from Google and Meta are examples of large language models. They are trained using tens of trillions of words of text — as much as all the books humanity has produced — to imitate human linguistic behavior. They result from billions of trillions of small random perturbations in the training process. They are not designed or programmed in any meaningful sense. They do not follow rules. Like chess programs, they may be pursuing objectives, but we have no idea what those objectives are. To get LLMs to behave themselves, OpenAI employs thousands of human trainers to say the equivalent of "Bad dog!" whenever the systems misbehave. And misbehave they still do — advising on ways to commit suicide or build biological weapons, practicing law and medicine without a license, and committing dozens of other categories of transgressions. LLMs are also notorious for "hallucinating" — generating completely false answers, often supported by fictitious citations — because their training has no connection to an outside world and the truth of assertions about it.

The tech companies' lobbyists will complain that their wonderful systems cannot possibly meet the required safety criteria. So be it. We would never accept such an excuse from pharmaceutical manufacturers or from builders of nuclear power stations, and we should not accept it from purveyors of AI systems.

The core problem is that neither OpenAI nor anyone else has any real idea how LLMs work. I asked Sébastien Bubeck, lead author of the paper "Sparks of Artificial General Intelligence: Early experiments with GPT-4," whether GPT-4 had developed its own internal goals. The answer? "We have no idea."

Reasonable people might suggest that it's irresponsible to deploy — on a global scale — a system that operates according to unknown internal principles, that shows "sparks" of general-purpose intelligence, and that may or may not be pursuing its own internal goals. At the moment, there are technical reasons to suppose that GPT-4 is limited in its ability to form and execute complex plans, but dozens of research groups are exploring ideas for overcoming this and other limitations.

Just as with the impending arrival of a superior alien civilization, it is imperative that governments cooperate on the regulation of AI. It's in no country's interest for any country to develop and release AI systems that humans cannot control. This is the question underlying the open letter: How do we retain power over entities more powerful than us, forever?

Stuart Russell is a professor of computer science at the University of California, Berkeley; director of the Center for Human-Compatible Artificial Intelligence; and director of the Kavli Center for Ethics, Science, and the Public.