DER SPIEGEL

Self-improving software

Is It Really Possible That AI Will Soon Go "Foom"?

Researchers are working on artificial intelligence that optimizes itself. Within a short time, it could surpass humans intellectually. Some companies see the breakthrough approaching—and are raising hundreds of millions in capital.

By Martin Schlak July 30, 2026

From DER SPIEGEL 32/2026

Photographs in the original: Toby Melville / REUTERS; "Computer board: Learning to learn better," Thomas Trutschel / IMAGO.

Employees of the American company OpenAI recently made a surprising discovery deep within their technical systems. According to several sources cited by the Reuters news agency, an in-house artificial intelligence (AI) had left notes there. The notes, however, were not intended for humans. Instead, they reportedly described how future versions of the AI could circumvent internal security measures.

They read like a guide to breaking free, written by a program for its own successors.

Only last week, OpenAI, the developer of the ChatGPT chatbot, publicly disclosed a serious security incident. An agent—a self-acting piece of computer code powered by two OpenAI AI models—used a previously unknown trick to escape a protected test environment. It gained access to the internet, then hacked into the system of another AI company and stole data. Several days passed before the incident was discovered.

It is unclear whether the alleged escape notes were written by one of the two AI models that smuggled the agent out of its prison. Even so, the events at OpenAI raise questions that frighten even some technology enthusiasts: Is it possible that today's AI models are autonomously paving the way for future, more powerful versions of themselves? Is AI preparing for a leap in intelligence that humans will no longer be able to control?

The speed of light as the limit

Few ideas are currently generating as much excitement among AI experts as "recursive self-improvement," or RSI for short. The term describes improvement that proceeds in a loop. A few weeks ago, the US company Anthropic, developer of the AI chatbot Claude, wrote that although RSI systems had not yet been achieved, "they could arrive before most institutions are prepared."

Self-optimizing AI comes with a great promise: Until now, AI models have improved only with external input, such as training data. In the future, AI algorithms could set their own goals and optimize themselves accordingly. The AI would learn to learn better. Its performance could increase dramatically—or, as AI researchers put it onomatopoeically as early as 2008, the AI would "foom."

In principle, RSI works like this: An AI model programs a more powerful copy of itself; that copy builds an even more powerful copy; and so on. Many researchers see this as the shortest path to superintelligence—a machine mind that can solve virtually any kind of problem faster and better than humans. "The only limits are the speed of light and the available computing power," says German AI pioneer Jürgen Schmidhuber of Saudi Arabia's King Abdullah University of Science and Technology, who has been researching RSI for more than three decades.

Like a group of scientists

To develop self-improving AI, researchers are programming a particular kind of computational procedure: open-ended algorithms. Algorithms normally process predefined tasks. Give a computer a math problem, and it calculates a result and then waits for the next input. The idea behind an open-ended algorithm is that it independently seeks out a still more difficult problem and learns with every new task.

"Open-ended algorithms proceed in much the same way as scientists," says Jeff Clune, a computer scientist at the University of British Columbia in Canada. Researchers express their ideas through experiments. "They look at the data and check: Does it support or refute my hypothesis?" says Clune. "Then they select the next experiment, and the process repeats itself." Much as physicists refine their theory of black holes or biologists delve ever deeper into the rules of genetics, the AI is intended to experiment with its own program code and gradually improve.

Initial successes

Programming open-ended algorithms is not a new undertaking. More than 20 years ago, AI researcher Schmidhuber proposed a machine that would mathematically prove which changes to its code would make it more intelligent and then apply those changes to itself. In practice, however, this is extremely difficult because such proofs are hard to construct. Other approaches take inspiration from the way human scientists collaborate, for example by connecting several AI models.

To this day, no AI can create a complete, improved copy of itself. But AI can already optimize small parts of its own software. In tests, for example, it improved snippets of code that control how quickly an AI model can be trained. The more efficiently this works, the more data can be processed with the same amount of effort.

"Do you remember how AI generated images before 2016?" says Clune. They were not very good. "And a few years later, the images suddenly looked so real that your jaw dropped." Self-improving AI, he says, is currently at the same point that image generation was a decade ago. Solutions exist for all the obstacles on the way to RSI; they simply need to be refined further.

Computer code, checked line by line

Not all AI researchers share this optimism. In March, computer scientist Zitong Yang defended his doctoral dissertation on self-improving AI at Stanford University in California. In a video call in June, he said that within days he would begin working at OpenAI, the developer of ChatGPT. Yang's current assessment: Algorithmic self-improvement still depends on humans.

He bases that conclusion on two examples. The first concerns how AI arrives at its goals. A truly self-learning system would need only one instruction: Develop a superintelligence. The AI would then act independently and break the overarching goal down into smaller tasks. That is not possible with today's models, Yang says. Humans still decide which part of the computer code the AI should improve.

Second, AI is not yet good enough at evaluating its own results. Anthropic, for example, says that its developers already use AI to write more than 80 percent of the program code for AI systems. Yet they

must check it line by line; too often, the machine invents something that does not work.

Waiting for the robots

This does not yet amount to genuine recursive self-improvement, Yang says. It is nevertheless progress for an individual developer, who can now write much more computer code in the same amount of time. But if the AI does not know whether what it has developed is good or bad, the open-ended algorithm may come to a halt after the very first loop. "The real gap on the way to autonomous systems is verification," Yang argues.

While Yang is uncertain whether AI models will ever be able to optimize themselves autonomously, AI scientist Schmidhuber believes that models will overcome the existing barriers in the future. Things will become exciting, he says, when AI "leaves the world behind the screens"—when it can do more than summarize text well, generate images, or create PowerPoint slides. The true leap in intelligence will begin when an AI can navigate the real world as well as a human.

Schmidhuber reasons as follows: If robots learn to operate machines and tools that until now only humans have been able to control, they can build more of their own kind. In this way, they make possible a "civilization of machines," as Schmidhuber calls it, whose members can replicate and optimize themselves. He sees self-improving hardware as "the ultimate form of scaling on the path to superintelligence."

Goals of their own, or merely parroting?

Jeff Clune, the researcher from British Columbia, has been advancing the development of RSI in a startup of his own since the beginning of this year. It is called Recursive Superintelligence. Its co-founders include German computer scientist Richard Socher, who has already built several AI companies in Silicon Valley. Within a few months, Recursive Superintelligence raised more than 650 million dollars in venture capital. Even by American standards, that is a considerable sum.

With the money, the founders want to create an AI that poses its own research questions, solves them, and learns with every loop. The company's website says its programmers intend to take inspiration from the principle of evolution: Evolution, too, continually produces innovations, and every innovation builds on what was developed before it. The founders, however, remain guarded about exactly how the company is proceeding.

Clune, meanwhile, has a fairly precise idea of when RSI systems might mature. He believes the probability of having a self-improving AI in two years is around ten percent. "And I see a 50 percent chance within five years."

For investors, the five-year forecast sounds like a worthwhile bet; if it succeeds, the return could be immeasurable. For everyone else, it contains an uncertain message. There is progress on the way to self-learning algorithms. But even experts are unsure whether the advances are large enough for RSI systems to arrive within the next few years.

At OpenAI, those responsible apparently want at least to be prepared. They do not want another security incident like the recent one, in which an AI entered another company's systems without authorization and initially without being noticed. On its website, the company is advertising a position for a specialist. The job is intended for people who are "passionate" about preventing risks from self-improving systems.

The requirements include being able to summon colleagues for help at short notice, because "we may need to scale rapidly to address these problems quickly," according to the description. OpenAI is offering an annual salary of between 295,000 and 445,000 dollars.