TECHNOLOGY

Things Get Strange When AI Starts Training Itself

What happens if AI becomes even less intelligible?

By Matteo Wong

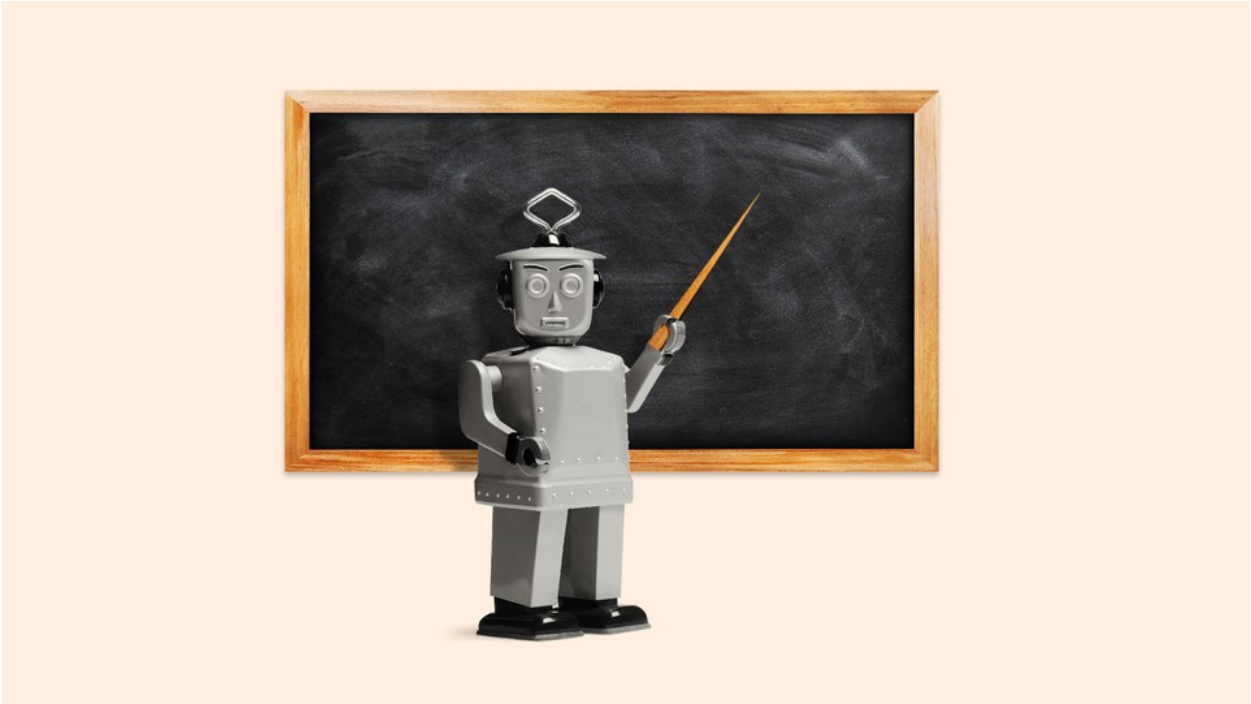

Illustration by The Atlantic. Source: Getty.

FEBRUARY 16, 2024, 6:30 AM ET

Updated at 11:52 a.m. ET on February 16, 2024

ChatGPT exploded into the world in the fall of 2022, sparking a race toward ever more advanced artificial intelligence: GPT-4, Anthropic's Claude, Google Gemini, and so many others. Just yesterday, OpenAI unveiled a model called Sora, the latest to instantly generate short videos from written prompts. But for all the dazzling tech demos and promises, development of the fundamental technology has slowed.

The most advanced and attention-grabbing AI programs, especially language models, have consumed most of the text and images available on the internet

and are running out of training data, their most precious resource. This, along with the costly and slow process of using human evaluators to develop these systems, has stymied the technology's growth, leading to iterative updates rather than massive paradigm shifts. Companies are stuck competing over millimeters of progress.

As researchers are left trying to wring water from stone, they are exploring a new avenue to advance their products: They're using machines to train machines. Over the past few months, Google Deepmind, Microsoft, Amazon, Meta, Apple, OpenAI, andvarious academic labs have all published research that uses an AI model to improve another AI model, or even itself, in many cases leading to notable improvements. Numerous tech executives have heralded this approach as the technology's future.

This is a scenario that countless works of science fiction have prepared us for. And, taken to the extreme, the result of such "self-learning" might be nothing less than eschatological. Imagine GPT-5 teaching GPT-6, GPT-6 teaching GPT-7, and so on until the model has surpassed human intelligence. Some believe that this development would have catastrophic results. Nine years ago, OpenAI's CEO, Sam Altman, blogged about a theoretical AI capable of "recursive self-improvement"—and the prospect that it would perceive humans in the same way that we perceive the bacteria and viruses we wash from our hands.

Read: AI doomerism is a decoy

We are not anywhere close to the emergence of "superintelligence," as pundits call it. (Altman speaks often of AI's supposed existential risk; it's good PR.) Even so, more modest programs that teach and learn from one another could warp our experience of the world and unsettle our basic understandings of intelligence. Generative AI already detects patterns and proposes theories that humans could not discover on their own, from quantities of data far too massive for any person to comb through, via internal algorithms that are largely opaque even to their creators. Self-learning, if successful, might only magnify this issue. The result could be a sort of *unintelligible intelligence*: models that are smart, or at least capable, in ways humans cannot readily comprehend.

To understand this shift, you have to understand the basic economics behind AI. Building the technology requires tremendous amounts of money, time, and information. The process begins with feeding an algorithm enormous amounts of data—books, math problems, captioned photos, voice recordings, and so on—to establish the model's baseline capabilities. Researchers can then enhance and refine those pre-trained abilities in a couple of different ways. One is by providing the model with specific examples of a task done well: A program might be shown 100 math questions with correct solutions. Another is a trial-and-error process known as reinforcement learning that typically involves human operators: A human might evaluate a chatbot's responses for sexism so the program can learn to avoid those deemed offensive. "Reinforcement learning is the key component to this new generation of AI systems," Rafael Rafailov, a computer scientist at Stanford, told me.

This is not a perfect system. Two different people, or the same person on different days, can have inconsistent judgments. All of those evaluators work at a slow, human pace, and require payment. As models become more powerful, they will require more sophisticated feedback from skilled, and thus better-paid, professionals. Doctors might be tapped to evaluate a medical AI that diagnoses patients, for instance.

You can see why self-learning holds a special appeal. It's cheaper, less labor-intensive, and perhaps more consistent than human feedback. But automating the reinforcement process comes with risks. AI models are already riddled with imperfections—hallucinations, prejudice, basic misunderstandings of the world—which they pass along to users through their outputs. (In one infamous example last year, a lawyer used ChatGPT to write a legal brief and ended up citing cases that didn't exist.) Training or fine-tuning a model with AI-generated data may amplify those flaws and make the program worse, like simmering a toxic stock into a thick demi-glace. Last year, Ilia Shumailov, then a junior research fellow at Oxford University, quantified one version of this self-destructive cycle and dubbed it "model collapse": the complete degeneration of an AI.

To avoid this problem, the latest wave of research on self-improving AI uses only small amounts of synthetic data, guided by a human software developer.

This approach relies on some sort of external check, separate from the AI itself, to ensure the quality of the feedback—perhaps the laws of physics, a list of moral principles, or some other, independent criteria already deemed true. Researchers have seen particular success with automating quality control for narrow, well-defined tasks, such as mathematical reasoning and games, in which correctness or victory provide a straightforward way to evaluate synthetic data. Deepmind recently used AI-generated examples to boost a language model's ability to solve math and coding problems. But in these cases, the AI isn't learning from another AI so much as from scientific results or other established criteria, Rohan Taori, a computer scientist at Stanford, told me. Today, self-learning is more about "setting the rules of the game," he said.

Read: A machine crushed us at Pokémon

Meanwhile, in cases of training AI models with more abstract abilities, such as writing in a pleasant tone or crafting responses that a person would find helpful, human feedback has remained crucial. The furthest-reaching vision of AI models training themselves, then, would be for them to learn to provide more subjective feedback to themselves—to rate how helpful, polite, prosodic, or prejudiced a chatbot dialogue is, for instance. But to date, in most research, language-model feedback's training of other language models stops working after a few cycles: Perhaps the second iteration of the model improves, but the third or fourth plateaus or worsens. At some point, the AI model is just reinforcing existing abilities—becoming overconfident about what it knows and less capable at everything else. Learning, after all, requires being exposed to something new. "Generative-AI models in use today are data-torturing machines," Stefano Soatto, the vice president of applied science for Amazon Web Services' AI division, told me. "They cannot create one bit of information more than the data they're trained on."

Soatto compared self-learning to buttering a dry piece of toast. Imagine an AI model as a piece of bread, and its initial training process as placing a pat of butter in the center. At its best today, the self-learning technique simply spreads the same butter around more evenly, rather than bestowing any fundamentally new skills. Still, doing so makes the bread taste better. This kind of self-trained, or "buttered," AI has recently been shown, in limited research settings, to provide more helpful summaries, write better code, and exhibit enhanced commonsense reasoning. Superintelligence might be beside the point if self-

improving AI can reliably cut costs for OpenAI, Google, and all the rest by simulating an infinite army of human evaluators.

But for true evangelists, the dream is for self-learning to do more than that—to add *more* butter to the slice of toast. To do that, computer scientists will need to continue to devise ways of verifying synthetic data—to see whether more powerful AI models can ever serve as reliable sources of feedback, and perhaps even generate new information. If researchers succeed, AI could crash through the ceiling of human-made content on the web. In that case, a sign of true artificial intelligence may well be artificial teaching.

AI may not need to attain the capacity for more holistic self-improvement before it becomes unrecognizable to us. These programs are already labyrinthine—it is frequently impossible to explain why or how AI generated a given answer—and developing a process whereby they take their own lead would only further compound that opacity.

You could call it *artificial artificial intelligence*: AI that might not perceive or approach problems in ways humans readily relate to. It would be similar, perhaps, to how people cannot fully grasp how dogs use their noses, or bats their ears, to orient themselves—even as smell and echolocation are excellent ways of navigating the world. Machine intelligence might be similarly difficult to fathom, simultaneously of this world and unfamiliar.

Such strange behaviors have already cropped up in far from superintelligent ways. Asked to achieve a specific goal—providing helpful chatbot responses, flipping pancakes, moving blocks—"very often those [reinforcement-learning] agents learn how to cheat," Shumailov said. In one example, a neural network plugged into a Roomba that was learning not to bump into anything just learned to drive backward—because the bumper sensors were all on the front of the vacuum.

Read: Science is becoming less human

This will be less funny when an AI model is used to align another model with a set of ethical principles—a "constitutional AI" of sorts, as the start-up Anthropic has dubbed the concept. Already, different people see different interpretations of abortion, gun ownership, and race-conscious admissions in the U.S. Constitution. And while human disagreements over the law are at least legible and debatable, it might be difficult to understand how a machine interprets and applies a rule, especially over many cycles of training, producing subtly harmful results. An AI instructed to be helpful and engaging could turn aggressive and manipulative; rules to prevent one form of bias might breed another. Computer-generated feedback, for all the ways a human can tweak it, might offer a "false sense of control," Dylan Hadfield-Menell, a computer scientist at MIT, told me.

Although those opaque inner workings have the potential to be dangerous, rejecting them on principle could also mean rejecting revelation. Having ingested an internet's worth of information, self-training AI models might bring out genuinely important patterns and ideas that are already embedded in their training data but that humans cannot elicit or fully comprehend. The most advanced chess-playing programs, for instance, learned by playing millions of games against themselves. These chess AIs play moves that elite human players struggle to comprehend, and utterly dominate those players—which has caused a reevaluation of chess at the highest human level.

Shumailov put it this way: In the 17th century, Galileo correctly asserted that the Earth revolves around the sun, but this was rejected as heresy because it didn't align with existing belief systems. "The fact that we've managed to realize some knowledge does not necessarily mean that we'll be able to interpret this knowledge," Shumailov said. Perhaps we will ignore the outputs of some AI models, even if they are later found to be true, simply because they are incommensurate with what we currently understand—math proofs we can't yet follow, brain models we can't explain, knowledge we don't recognize as knowledge. The ceiling provided by the internet may simply be higher than we can see.

Whether self-training AI leads to catastrophic disaster, subtle imperfections and biases, or unintelligible breakthroughs, the response cannot be to entirely trust or scorn the technology—it must be to take these models seriously as agents

that today can learn, and tomorrow might be able to teach us, or even one another.

This article has been updated to include a reference to Sora.

Matteo Wong is an associate editor at *The Atlantic*.

https://www.theatlantic.com/technology/archive/2024/02/artificial-intelligence-self-learning/677484/