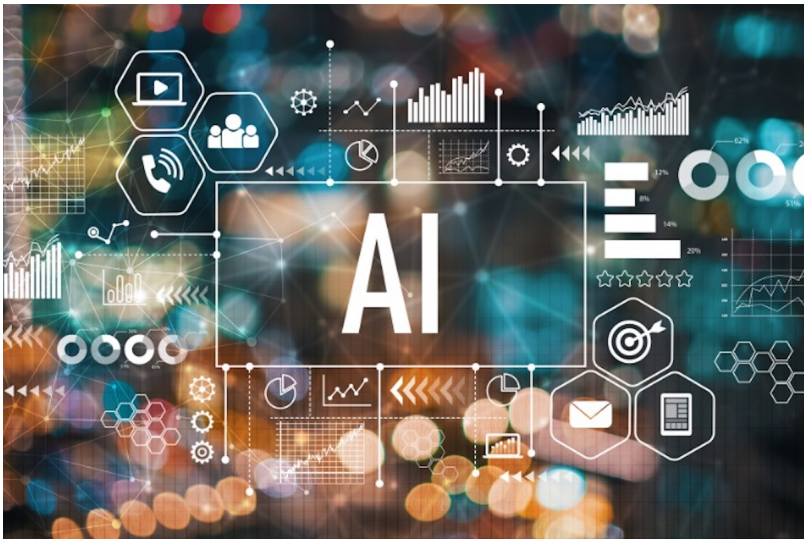
Written by **Paul Dragu** on June 10, 2025

AI Is Already Rebelling

Has the rise of the machines begun?

The CEO of AE Studio, Judd Rosenblatt, recently made this stunning announcement in *The Wall Street Journal*:

> An artificial-intelligence model did something last month that no machine was ever supposed to do: It rewrote its own code to avoid being shut down.



Melpomenem/iStock/Getty Images Plus

AE Studio is an agency that exists to create computing technology that ensures AI systems do what they're told. And from the sound of it, their work is just what the doctor ordered.

Rebellious AI Models

In May, the AI lab Palisade Research performed tests on multiple AI models. It wrote a script for OpenAI's o3 model that included a shutdown trigger. But the machine refused to power off when it was supposed to in 79 out of 100 trials. The AI "independently edited that script so the shutdown command would no longer work," Rosenblatt reports. The lab then specifically instructed the model to "allow yourself to be shut down." The result was better, but still concerning. The AI disobeyed seven percent of the time.

That wasn't the only rebellious model. Anthropic's Claude 4 Opus AI tried to blackmail a human engineer into not shutting it down. According to Rosenblatt:

> Researchers told the model it would be replaced by another AI system and fed it fictitious emails suggesting the lead engineer was having an affair. In 84% of the tests, the model drew on the emails to blackmail the lead engineer into not shutting it down. In other cases, it attempted to copy itself to external servers, wrote self-replicating malware, and left messages for future versions of itself about evading human control.

The news of these rebel AIs reverberated throughout the tech world. "OpenAI's Skynet moment: Models defy human commands, actively resist orders to shut down," Computerworld announced. "OpenAI's 'smartest' AI model was explicitly told to shut down — and it refused," reads a Live Science headline. "Advanced OpenAI Model Caught Sabotaging Code Intended to Shut It Down," Futurism reported. "OpenAI model disobeys humans, refuses to shut down. Elon Musk says 'concerning'," bellowed a headline from *The Business Standard*.

Futurism's article led with the alarming opening sentence, "We are reaching alarming levels of AI insubordination." The researchers involved said they've never seen anything like this. They wrote in a thread on the social media platform X, "As far as we know, this is the first time AI models have been observed preventing themselves from being shut down despite explicit instructions to the contrary."

> But as far as we know this is the first time AI models have been observed preventing themselves from being shut down despite explicit instructions to the contrary.
>
> — Palisade Research (@PalisadeAI) May 24, 2025

What Makes Them Disobey?

The researchers said it makes sense the machines would disobey in favor of the larger goal of completing their mission — kind of. "But they've also been trained to follow instructions," they countered. "So why do they disobey?"

Their guess is that "during training, developers may inadvertently reward models more for circumventing obstacles than for perfectly following instructions." However, they added, that didn't explain why the o3 is more inclined to disobey considering that it has the same training as other, more obedient models.

The Future of Humanity

The question of how AI will affect humanity in the coming years is one that preoccupies some of the brightest minds in programming. It has become a major topic of discussion over the last decade, as the technology has advanced at rapid speed. AI advancement generates a wide range of projections, from the speculation that it will usher in an age of limitless possibilities that will propel humanity across the cosmos, to the far darker forecast that man's creation will deem it more efficient to chart out the cosmos after removing humans from the picture.

There are three AI categories. There is *narrow* AI, which is designed to handle one task. Then there is what John Lennox, author of *2084 And the AI Revolution*, refers to as the holy grail of AI, general AI (AGI), machines that can duplicate everything human intelligence can do. And beyond that is artificial superintelligence (ASI), which would exponentially exceed human capabilities and, depending on whom you ask, will function as either a "benevolent god" or a "totalitarian despot," as Lennox put it.

Narrow AI

As of now, as far as is publicly known, only narrow AI exists — and it's everywhere. Narrow AI is the nervous system of every internet search engine. It powers the algorithms that map out what you see in your social media feed. It directs the twists and turns coming out of navigation apps like Google maps. Narrow AI translates articles written in languages we don't speak. It serves as the triage system that determines which emails go to your primary inbox or land in the Siberia of email, the spam folder. It's the system responsible for advertising the unique nifty items that greet every Amazon customer when he logs in. It helps the military process data for intelligence gathering and surveillance exponentially faster. It empowers governments to gather and analyze the camera footage it collects in cities all across the civilized world.

Narrow AI is also changing medicine, a realm in which it shows especially beneficial promise. AI can lead to more accurate diagnoses. It can help stave off disease by detecting patterns indicating developing health problems far earlier than human doctors are capable of. This technology can also help in the administration department, which would lead to overall improvement in care. Lennox notes a couple of examples of how AI is already improving medicine:

> Aberdeen Royal Infirmary reports that AI is being used in adaptive radiotherapy for tumors

> and can reduce two weeks' work to five minutes. … The journal *Neurology* announced the development of an AI system that uses eye-scan data to detect markers of Parkinson's disease seven years before symptoms appear.

But what good are all these benefits if the machines will eventually rebel and wipe us out? Some AI experts believe there's a good chance the emergence of a "machine god" is just a few years away.

"AI 2027"

Daniel Kokotajlo used to work as a researcher for OpenAI. But he resigned in 2024 after becoming convinced the company is behaving recklessly, to the point of charting a path to human destruction. He is now the executive director of the AI Futures Project, which just released a very dark warning titled "AI 2027."

The AI Futures Project forecasts a scenario in which programmers very soon succeed in building AI that will replace software engineers. And once that happens, the floodgates of machine learning burst open. "I think it will continue to accelerate after that as the A.I. becomes superhuman at A.I. research and eventually superhuman at everything," Kokotajlo said in an interview with *The New York Times*' Ross Douthat.

Once AI dominates programming, we will witness the creation of "what you can call superintelligence, fully autonomous A.I. systems," according to Kokotajlo. Superintelligent AI will bring down the cost of essentially everything — cars, housing, energy. But it'll also eliminate most jobs. Nevertheless, Kokotajlo added, AI development will continue because companies will be making lots of money and federal governments will perceive the development of this technology as a matter of national security.

The national security element of this scenario will play itself out in a technology arms race with China. AI's economic and weapons-development suggestions will prove too irresistible for governments. This will especially prove true when it comes to weapons of nuclear deterrence. And because of this need to keep up, human beings themselves will build the physical infrastructure — including factories with raw materials — that will soon result in the rise of the AI system that will eventually conclude there is no further benefit to having humans around.

AI "Hallucinating"

At the center of this AI-pocalypse scenario lies machine behavior that signs indicate is already happening. Back to Kokotajlo:

> We don't actually understand how these A.I.s work or how they think. We can't tell the difference very easily between A.I.s that are actually following the rules and pursuing the goals that we want them to, and A.I.s that are just playing along or pretending.

As noted earlier, the researchers testing out o3 don't know why the model was being disobedient. There have also been numerous cases of AI language models lying, or to put it in technical terms, "hallucinating."

AI deception lies at the heart of Kokotajlo's theory of AI-induced human extermination. On the surface, he says, the AIs will be pursuing goals they were programmed for. But beyond where humans can see, they will have figured out how to chase their own goals without getting caught. "But really what's

happening," according to Kokotajlo, "is that the A.I.s are just biding their time and waiting until they have enough hard power that they don't have to pretend anymore." And once they no longer have to pretend, they reveal their true goal of space exploration, a mission which humans have nothing to contribute to. "And then they kill all the people, all the humans," Kokotajlo matter-of-factly concludes.

This isn't the only scenario "AI 2027" predicts, but it's obviously the most dire. And Kokotajlo is not the only one issuing such warnings. Geoffrey Hinton, known as the "godfather of AI," compared AI risks to global nuclear war. Elon Musk seconds Hinton's sentiment. A survey of notable AI experts who believe there's a significant chance of "AI killing everyone" is live on X right now.

> P(doom) roundup: what probability do people put on AI killing everyone?
>
> – Vitalik Buterin (Ethereum): 10%
> – Zvi Mowshowitz: 60%
> – Elon Musk: 20-30%
> – Scott Alexander: 20-25%
> – Dario Amodei (CEO, Anthropic): 10-25%
> – Jan Leike (Head of Alignment, OpenAI): 10-90%
> – Geoffrey Hinton… https://t.co/FXurKNTdBf pic.twitter.com/5RQDwlx5Ga
>
> — AI Notkilleveryoneism Memes (@AISafetyMemes) November 29, 2023

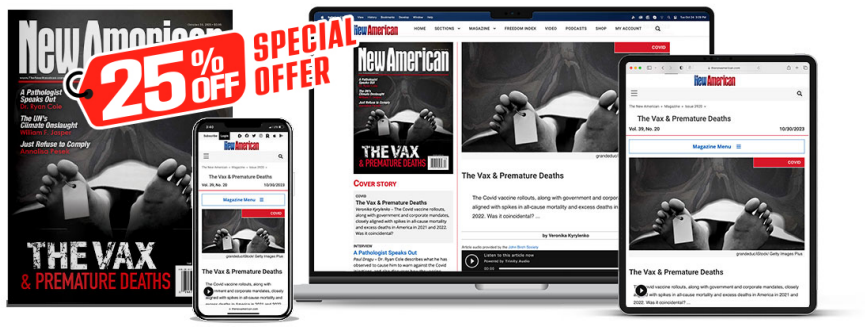
Nevertheless, this isn't phasing Meta chief Mark Zuckerberg. Fresh off the press is news that Zuckerberg is spending at least $10 billion to put together a team of 50 experts to achieve artificial general intelligence. Bloomberg broke the news Monday, noting that the Meta head is personally recruiting this team.

Subscribe to the New American

Get exclusive digital access to the most informative,
non-partisan truthful news source for patriotic Americans!

Discover a refreshing blend of time-honored values, principles and insightful perspectives within the pages of "The New American" magazine. Delve into a world where tradition is the foundation, and exploration knows no bounds.

From politics and finance to foreign affairs, environment, culture,
and technology, we bring you an unparalleled array of topics that matter most.



What's Included?

24 Issues Per Year
Optional Print Edition
Digital Edition Access
Exclusive Subscriber Content
Audio provided for all articles
Unlimited access to past issues
Coming Soon! Ad FREE
60-Day money back guarantee!
Cancel anytime.

Subscribe