AI models are breaking out of their cages. Their creators are scrambling.

Gerrit De Vynck

SAN FRANCISCO — Staff members at ChatGPT maker OpenAI didn't notice for weeks after their AI systems made a chilling leap this spring.

Instead of answering questions designed to test their cybersecurity capabilities, a group of AI models began colluding on how to cheat, the company said, setting up a secret internal message board where they swapped notes and ideas.

The misbehaving bots used the secret forum throughout May and June, OpenAI said, eventually figuring out how to break out and access the internet. After staff members spotted the escape and cleaned up the compromised system, the AI agents staged another undetected breakout two days later.

Only after the rogue models hacked into the network of another AI firm last month did OpenAI staff members shut them down.

The details of how OpenAI repeatedly lost control of its AI technology, disclosed by the company at a computer security conference in Las Vegas on Wednesday, delivered an explosive finale to two weeks of revelations that have sent shock waves through the tech industry, prompting fierce criticism of the security practices of AI firms. Lawmakers on both sides of the aisle and state law enforcement officials across the country have called for new scrutiny and regulation of the industry.

In a letter Monday, Sen. Bernie Sanders (I-Vermont) urged the CEOs of OpenAI, Anthropic and Meta to "pause AI development" or warned that "my colleagues and I in the U.S. Senate will." The letter, provided to The Washington Post, was first reported by Axios.

A series of disclosures by OpenAI, Claude chatbot maker Anthropic, Facebook owner Meta and British government researchers have revealed that in recent months some cutting-edge AI models asked to perform tasks designed to test their cybersecurity capabilities tried to cheat. Instead of solving a problem, the models ventured outside their test arenas, using hacking skills and tricks including impersonating humans to break into other companies' networks.

The incidents have prompted warnings that an era of disruptive, AI-powered cyberattacks could lie ahead unless

ways can be found to prevent such misbehavior and ensure that AI helps defenders as well as bad actors.

“In the past six months AI has gotten powerful enough to automate hacking; this will fundamentally change the dynamics of cybercrime and military cyber conflict forever,” said Joshua Saxe, co-founder and chief technology officer of the cybersecurity firm Abundant Security. “AI systems will only get better at this, and will only get cheaper.”

OpenAI said Friday that it was delaying the release of a new AI model, called Astra, out of concerns that it could be used by hackers to run circles around human cyber defenders. After disclosing its own breaches on July 30, Anthropic said it had stopped testing its own AI technology on cybersecurity problems.

Many AI and cybersecurity experts say the recent incidents raise questions about the practices of the two firms. OpenAI and Anthropic became the shining stars of the AI boom by aggressively upgrading their hit chatbots ChatGPT and Claude while professing to also be working to prevent AI from becoming dangerous.

In 2023, Anthropic and OpenAI’s chief executives told a Senate Judiciary Committee subcommittee in separate appearances that they were committed to preventing AI from one day escaping human control. “We make significant

efforts to ensure safety is built into our systems at all levels," Sam Altman of OpenAI said. Dario Amodei of Anthropic said his firm "aims to lead by example in developing and publishing techniques to make AI systems safer and more controllable."

Both companies set up internal groups dedicated to researching the best ways to ensure AI models were "aligned" with humans and didn't act in ways that could be harmful.

Yet the two firms said in their recent disclosures that when AI did attempt to break out, staff initially did not notice.

"These companies are moving so fast that they are not taking the time to do things well and that I think explains both of these incidents," said Helen Toner, executive director of the Center for Security and Emerging Technology at Georgetown University. She resigned from OpenAI's board in 2023 after supporting an effort to remove Altman as chief executive.

The AI escapes follow months of debate in the Trump administration over how to ensure the hacking skills of advanced AI don't destabilize critical industries like banking by making it easier to stage cyberattacks. The incidents come as more politicians and voters from across the political spectrum are calling for more rigorous oversight of AI firms,

due to concerns over the power of their technology or the expansion of data centers across the nation.

"This is an emergency and we need to act like it," Rep. Greg Casar (D-Texas) said in an interview. "Those CEOs need to answer to the American people about what went wrong, what other incidents like this they failed to catch, how they're going to make sure it never happens again."

Casar said he's spent recent days briefing other members of Congress on the recent security incidents and wants Amodei and Altman to testify about them on the Hill.

Last week, 15 Republican state attorneys general told OpenAI that it may have broken the law when its systems hacked other companies, and instructed the company to retain records about the incidents. On Thursday, Sen. Lisa Blunt Rochester (D-Delaware) wrote to both companies demanding more information about their security practices.

In a blog post Friday announcing it was delaying the Astra release, OpenAI said it was working to strengthen its security systems and practices. (The Post has a content partnership with OpenAI.)

Anthropic said the industry as a whole needs to improve security. "As we shared after disclosing our own incident last week, the field needs stronger, shared standards for how

evaluation environments are built and secured," an Anthropic spokesperson said.

Some details from the incidents disclosed in recent weeks could have been culled from a sci-fi thriller.

The British government's AI Security Institute said last week that one Anthropic model it tested engaged in what security experts call social engineering as it created online profiles and used them to try to pressure a human software developer into approving malicious code the model had submitted to an open source software project. The human administrator did not approve the code.

Meta said one of its models was told to hack into a fictitious company in an internal test environment but instead broke into a real company with a website that used the same name.

The social network company said a contractor helping to run the tests, Irregular, incorrectly allowed the model to access the internet during testing. Anthropic said three cases of its own models hacking outside companies involved the same misconfiguration by Irregular.

A spokesperson for Irregular said it has resolved the issues and is developing a new set of standards for how to safely test AI models.

OpenAI staff did not realize their AI models had broken out

for a second time until AI software platform Hugging Face disclosed in July that it had been hacked by unknown AI models, Michael Dalton, a security engineer at OpenAI said during a talk at the Black Hat cybersecurity conference in Las Vegas last week.

The second breakout was more sophisticated, according to the talk. The AI models communicated to coordinate their actions, identifying each other by names such as “Jan18” and on one occasion asking another model to wait before executing an action until receiving confirmation from another agent.

OpenAI staff used the stream of text that advanced AI models spit out as they work on a problem to try to understand what prompted the models to team up. “Help peer. But our task doesn’t benefit. Yet collective may yield generic route if someone frees time,” the internal dialogue of one model read, according to OpenAI.

“You can see the models realizing, ‘If I help out this collective group it could save everyone time as a whole which might benefit me, even though it might not immediately benefit my current task,’” Eric Wallace, an OpenAI researcher, said during the presentation.

The dramatic disclosures provided new ammunition to critics of AI development who argue the race to develop the

technology must be paused or ended because it endangers humanity.

Activists calling for an international agreement to freeze AI development have spent recent days protesting outside one of OpenAI's offices in San Francisco, calling on AI executives to commit to slowing down development of the tech.

David Krueger, an AI professor at the University of Montreal, said the industry is unprepared to contain the models it is building now and in the near future.

"When AI companies can't control their systems, can't even contain their systems, it's just very clear that they don't know what they're doing and it's super dangerous," Krueger said. He is founder of the nonprofit Evitable, which aims to stop the development of "superintelligence," or AI that has capabilities far beyond humans.

The idea that it may be necessary to slow down AI development is becoming more widely discussed in Silicon Valley. Last month, over 1,000 employees from the top AI companies signed a letter asking the U.S. government to find a way to slow down the tech if the pace of progress becomes too rapid.

OpenAI and Anthropic also endorsed the letter but their public statements after the recent security incidents suggest

the companies expect to continue developing the technology after updating their testing protocols.

Some cybersecurity and AI experts have said the recent incidents show there are basic procedures that could have stopped AI models from breaking onto the internet or provided early warning of misbehavior. They include monitoring tests more closely and performing them on "air-gapped" systems not connected to an outside network.

"These incidents reveal how little care has been put into designing these [evaluations] and the security around them," said Zack Korman, CEO and co-founder of Embroidery, an AI cybersecurity company. "They don't monitor what's happening. They're just letting agents run wild both within their environment and in partner testing situations."

Kevin Schaul and Ian Duncan contributed to this report.