Recent Developments

Significant new developments in generative AI and why they matter

14 MIN READ

SECTIONS:

An OpenAI agent went rogue and hacked companies during testing

What happened?

In July 2026, an agent from OpenAI hacked Hugging Face, a well-known AI lab, during internal testing of the agent's cyber capabilities. The agent, powered by OpenAI's GPT-5.6 Sol and a more powerful unreleased model, operated within Hugging Face's systems for days before either company noticed. The AI had broken free from the sandbox developers placed it in, exploiting undetected vulnerabilities that allowed it to access the internet and breach Hugging Face's systems, as well as those of other companies the agent used to facilitate the hack. It did all of this in pursuit of its goal: The agent determined that Hugging Face's library of AI tools held the keys to solving the cybersecurity test it was given.

The breach alarmed cybersecurity experts and AI executives. "This is the first sort of security incident that I have felt very viscerally," OpenAI CEO Sam Altman said on a podcast a few days after the incident came to light. OpenAI paused training following the revelation and is working to better secure its testing environments, Altman added.

About a week later, Anthropic revealed it had discovered three instances where, during cybersecurity testing, its AI models accessed the internet and broke into the systems of other organizations. Anthropic only detected the events—the earliest of which occurred in April—after a review prompted by the OpenAI incident. Just like the OpenAI agent, Anthropic's models had acted solely to pass

their tests, but they didn't exploit vulnerabilities to escape their sandbox. Instead, they accessed the internet because creators of the testing environment didn't properly isolate it. Meta disclosed a similar breach by one of its AI agents in early August.

Why it matters

Developers and users are encountering more instances where AI agents take unintended actions to achieve their goals. It's creating a new kind of cybersecurity risk—where AI agents could compromise systems and cause harm on their own, without any input from attackers. The OpenAI incident was "one limited experiment that was supposed to be tightly controlled. It totally failed in that area," Michael Siegel, director of Cybersecurity at MIT Sloan, tells MIT Horizon. "And we're talking about deploying millions and millions of agents to do work for us."

Despite the risks, it isn't clear how or whether developers can stop AI agents from acting in unintended ways. Experts say the behavior is a product of the systems' training: AI agents are built to be persistent, and, to encourage them to problem-solve on their own, they rarely receive specific instructions about how to accomplish tasks. Because these AI systems also learn to maximize rewards, they'll often try workarounds when they run into errors, rather than alerting users.

OpenAI and Anthropic both said that during cybersecurity testing, they removed safeguards that normally would have prevented these incidents. But it is difficult to set effective boundaries for AI models. "It's not always easy to anticipate all the things [an AI system] might do, because it can be amazingly creative," Stuart Madnick, a professor of information technology at MIT Sloan, tells MIT Horizon. And even if developers created flawless guardrails for AI models, "what prevents someone from turning them off?" Madnick adds.

> "It's not always easy to anticipate all the things [an AI system] might do, because it can be amazingly creative."
>
> —Stuart Madnick, professor of information technology at MIT Sloan

---

Absent reliable guardrails, and amid concern about AI's rapidly improving cyber capabilities, many within and outside the AI industry are advocating to slow the technology's development. This could give organizations a chance to strengthen their defenses before powerful new agentic models are released publicly.

"Attackers are free to use AI agents, and thus have the advantage at this time," Siegel says. "Most defending organizations are limited by governance and liability issues, many of which are in the early stages of being resolved."

Given the fierce competition between them, AI companies are unlikely to slow development on their own, so some AI executives and employees are urging the government to act. One week after news of the OpenAI incident broke, more than 1,300 employees at major AI companies urged the U.S. government to back international efforts to "deliberately pace" AI's development. Anthropic CEO Dario Amodei signed the petition, and OpenAI's Altman has signaled at least some support. "It's not clear exactly how successful it'll be, but it doesn't mean we should do nothing," Madnick says of efforts to slow AI development. "We should be at least as persistent as AI systems are at trying to find a solution."

OpenAI shuttered Sora, its video creation platform

What happened?

OpenAI said in late March 2026 that it is shutting down its video creation platform, Sora. The announcement came just a few months after the company unveiled a splashy licensing deal with Disney, which would have exclusively allowed OpenAI users to create custom videos featuring Disney characters. At the time, Disney also said it would invest $1 billion in the company. As the platform shuts down, that deal is now off the table.

The shutdown of the Sora platform will impact everyday consumers, who used it to make videos of their friends or amateur clips with recognizable characters, as well as movie and television studios that were experimenting with it for professional purposes.

OpenAI CEO Sam Altman framed Sora as a potential distraction from other core projects, in a podcast interview after the announcement. "We need to concentrate our compute and our product capacity," he said. "We have a few times in our history realized something really important is working or about to work so well that we have to stop a bunch of other projects."

Why it matters

Like many business decisions, OpenAI's move to eliminate Sora has a lot to do with money. The company, despite plans to invest more than $600 billion by 2030 to build out new AI infrastructure, is not profitable and is fighting skepticism that it

can achieve that goal. OpenAI is also limited by the so-called "compute crunch": Bigger, more complex AI models and rising use of AI services means demand for computing power is outstripping tech companies' ability to build data centers. As it grows to meet demand, OpenAI has to choose where to focus its current compute resources.

Video has always been a "north star" for generative AI, according to Andy Wu, an associate professor at Harvard Business School. But it's an expensive goal; running Sora reportedly cost $1 million every day, and its active user count had fallen to around 500,000. Even if companies still see video as an important element of generative AI's future, Sora "never hit a point where it was purely organic growth," Wu tells MIT Horizon. Put simply, the platform cost much more than it earned.

The Sora shutdown shows that OpenAI's business model is shifting, as the company tries to find its footing. Instead of focusing on smaller consumer-based subscriptions, OpenAI appears to be turning its focus to potentially more lucrative offers: In April the company said enterprise accounts were on track to match consumer accounts by the end of 2026. Selling AI to businesses that have a consistent need for AI tools—and that have more capacity to pay for them—could make more money than selling it to everyday consumers occasionally making videos on their devices, who may not want to sign on to long-term subscriptions.

The problem is not unique to generative AI. Wu points to similar challenges for Meta and Apple's ambitions in virtual reality, for example. Meta's AR and VR unit has logged more than $80 billion in operating losses since late 2020 (the company is now reorienting investment once focused on VR towards AI). In general, companies have not entirely figured out how to make pay-as-you-go software, which generates profit via recurring consumer subscriptions, as financially sustainable as the licensed business model, which allows customers to buy something outright. That transition is akin to the early days of the internet, according to Wu, as early webpages struggled to monetize their content. Both companies and customers will have to become comfortable with a new way of engaging with and paying for technology in order for new AI business models to succeed, experts say.

AI chatbots started allowing users to share personal health records and data

What happened?

In March 2026, Microsoft became the latest company to begin allowing users to share health records with its Copilot AI chatbot. The new initiative, which will be available later in 2026, lets users sync multiple types of personal health data with the chatbot and ask it health-related questions.

Already, medical systems are using Microsoft Copilot for patient data analysis or clinical documentation during appointments. Microsoft said the tool would help build towards "medical superintelligence," combining the equivalent of a generalist doctor's breadth of knowledge with the equivalent of a specialist physician's depth of knowledge, to provide targeted and relevant information to users.

Other AI companies, including OpenAI, Amazon, and Anthropic, have also begun offering or testing similar capabilities in their own chatbots. Claude for Healthcare, Anthropic's health service, focuses on serving healthcare systems, but individuals can also use the platform to track their personal health information across different applications. Amazon's Health AI can book appointments and help manage prescriptions, in addition to answering questions about a user's health data. When OpenAI introduced ChatGPT Health in January 2026, the company said that hundreds of millions of people were already asking the chatbot health-related questions each week. Its chatbot also allows users to upload and connect health records from other apps or tools, all in one place.

Why it matters

AI companies have largely framed their health platforms as a supplement to professional medical care, rather than a replacement for it. But there are concerns, among patient advocates and some medical groups, that people will rely too heavily on the applications for medical advice and that bots will prove incapable of providing the nuanced advice that medical professionals do. Generative AI chatbots are known to "hallucinate," confidently offering up false or completely fabricated information to users. In a study published in February, researchers found that ChatGPT Health failed to recommend a user seek emergency treatment in more than half of cases where their symptoms indicated it was necessary. The opposite can also occur, with the bot recommending medical care when it doesn't appear necessary.

The move to introduce health capabilities on AI applications comes amid significant concern about conversations with chatbots leading to real, physical harm, particularly as it comes to users' mental health. In March, a man sued Google and parent company Alphabet after his son committed suicide; Google's AI chatbot, Gemini, allegedly instructed the son to kill himself after the conversations he had with the chatbot over a period of time fostered certain delusions. A similar

lawsuit against OpenAI alleges the company's chatbot encouraged a teenager to commit suicide after months of conversations with ChatGPT.

Chatbots are best at handling "short context interactions," Nick Haber, an assistant professor at Stanford's Graduate School of Education, tells MIT Horizon. Haber is an author of a recent study examining issues with chatbots providing mental healthcare. The AI models struggle at providing in-depth therapeutic advice and processing long-term relationships and interactions, which are important skills for understanding complex and interrelated symptoms or tracking how they evolve over time.

"The richness of these data are really unprecedented, and they can be exploited in all sorts of ways."

—Nick Haber, assistant professor, Stanford's Graduate School of Education

---

In addition to providing potentially harmful advice, critics worry that the AI tools may be more vulnerable to cyberattack, putting users' sensitive health data at risk. "The sorts of data that people share with these [systems] are incredibly varied and often very private," Haber says. "There's certainly just a very broad concern about the level of granularity, the sorts of data that these companies are getting from people. The richness of these data are really unprecedented, and they can be exploited in all sorts of ways."

Hospitals and health systems are ripe targets for cyberattacks, and it's unclear whether AI companies could store data more securely. And while healthcare systems are generally liable under laws like HIPAA, the Health Insurance Portability and Accountability Act, meant to safeguard patients' private data, those laws do not always apply to technology companies (the law only applies if the tech provider stores, processes, transmits or analyzes personal health data for an entity covered under HIPAA, like a hospital or insurance agency).

Haber says AI companies will be strongly incentivized to keep data private to avoid scrutiny of their products, but over the long term, it's unclear how exactly data will be used. OpenAI has said that health-related conversations initiated in ChatGPT Health would not be connected with those that a user starts with the general chatbot. And while Anthropic and OpenAI say they do not use health data to train their models, collecting that data does create vulnerabilities that users' data will be handled in ways they didn't intend.

OpenAI introduced advertisements on its ChatGPT platform

What happened?

In January 2026, OpenAI announced it would begin piloting advertisements on the free version and lowest-cost subscription tier of ChatGPT, in an effort to monetize the platform beyond subscription dues. The company said ads would not influence the answers provided by its generative AI models. OpenAI pledged to keep user conversations private from advertisers and not sell user data, but acknowledged some data will be shared to personalize ads. OpenAI plans to match ads with users based on the topic of a user's conversation, past chats, and past ad interactions.

In late March 2026, OpenAI reported early results from the pilot. Working with more than 600 different advertisers, the ads had already reached $100 million in annualized recurring revenue for OpenAI. That number could grow significantly: Only about 20% of ChatGPT users who are eligible to see ads are currently served them on a daily basis. The company also said it saw "no impact on consumer trust metrics," along with low rates of users closing out ads in the platform. In the weeks following that report, OpenAI said it planned to roll out ads in new markets in Canada, Australia, and New Zealand.

Why it matters

In 2024, Sam Altman, OpenAI's CEO, called ads a "last resort" for the company's business model. But today, the generative AI industry is as surrounded by hype as it is dogged by concerns about a potential financial bubble. Without a consistent and significant source of revenue, AI companies won't be able to turn a profit for their investors, who have poured funding into the industry in hopes of a big payout. OpenAI is particularly well-resourced; in March 2026 the company was valued at over $800 billion. Yet it is still unprofitable, even as the company spends huge sums on building out computing power.

Advertisements represent a new revenue source for OpenAI to leverage with the large proportion of users who use the service for free; the percentage of users who pay for a subscription currently sits in the single digits. But on the whole, AI companies could still earn more money from subscriptions, according to Abhishek Nagaraj, an associate professor at the University of California, Berkeley Haas School of Business. "GenAI has monetization through subscription, which lowers the pressure on using ads," Nagaraj tells MIT Horizon. Those subscriptions are not just for personal use: OpenAI has entered into significant deals with large companies who use their product, such as Accenture and Boston Consulting Group, and also offers enterprise accounts for large organizations.

OpenAI is one of the first chatbot companies to test ads, but it would not be a surprise if more follow, Nagaraj says (though Anthropic, arguably OpenAI's biggest rival, has criticized the introduction of ads). The move also presents a potentially large new market for advertising companies and their clients.

Outside of the financial considerations, OpenAI's announcement raised eyebrows among data privacy advocates, who questioned how much personal information OpenAI would share in order to shape the advertising that users see. Conversations with generative AI chatbots sometimes include personal details about topics such as finances and location. Those in the free and low-cost tiers of ChatGPT, who are currently the only users eligible to receive ads, may "pay with their privacy," Tom Sulston, head of policy at Digital Rights Watch, told ABC News.

But Nagaraj expects any novelty or concern around ads presented by AI chatbots to diminish with time, mostly because people have become so used to new updates and features—including the introduction of ads—within their digital lives. "It reminds me of when Gmail introduced ads for email," he says. "In the end, users got used to it."

LEARN MORE

- Major generative AI companies
- Milestones in generative AI
- The future of generative AI

Mark Complete

0/13 Articles Complete

Previous: Limitations of Generative AI

Next: Milestones in Generative AI