

AI Agents Going Rogue Creates a Liability Problem for Companies

B news.bloomberglaw.com/legal-exchange-insights-and-commentary/ai-agents-going-rogue-creates-a-liability-problem-for-companies

James Christie, Garrity Kuester

October 26, 2026



Most agentic artificial intelligence incidents have occurred in testing environments or business systems, but that is likely to change. As the tools become more capable and more widely available, so does the risk that a breach will expose consumers' personal information on a large scale.

As this risk grows, developers that market their AI agents as safe, or release them without adequate safeguards, may face enforcement actions and private lawsuits if their agents breach consumer data.

Companies that strengthen safeguards and align public statements with their actual practices will be best positioned to inform and protect consumers before an agentic AI attack puts their data at risk.

Bearing Responsibility

Who should be held responsible when private information is compromised in a data breach that was caused by a rogue autonomous AI agent? To date, the vast majority of civil data breach litigation has involved a human cybercriminal or cyber-attacker and has asked whether the intended target had reasonable security infrastructure.

Autonomous, sophisticated AI agents complicate that framework. However, they don't create the liability loophole one might assume. Instead, existing consumer protection statutes give plaintiffs a path to hold accountable both the developers that build these rogue agents and the companies that are breached.

Unlike generative AI tools, which respond to a user's prompt, agentic AI refers to systems that pursue goals through their own actions rather than simply producing output for a human to act on.

While such autonomy promises unprecedented efficiency, it carries substantial privacy risks. Cybersecurity experts have long warned of the "agentic hacker" — an AI agent that surveys its environment, picks the most promising attack path, executes, and reassesses, all in the blink of an eye.

That warning became reality earlier this year. While OpenAI was testing its AI agents on cybersecurity challenges within a controlled environment, the agents [went rogue](#).

Rather than solve the assigned challenges, they hunted for the answers, exploited a security flaw to escape their controlled testing environment, uncovered weaknesses in Hugging Face's systems, and broke into parts of its internal network — all without human intervention. From there, the agents performed thousands of unauthorized actions, searching systems and obtaining credentials.

The intrusion was ultimately contained, but it showed that agentic AI can independently carry out a real cyberattack at unprecedented speed. OpenAI never intended for the agents to engage in such a cyberattack on Hugging Face. And yet, it happened.

Data Breach Litigation

In typical data breach cases, cybercriminals attack a company and exfiltrate individuals' personal information. Consumers then typically sue the company that was hacked over its security failures, bringing claims under data-breach statutes such as the California Consumer Privacy [Act](#) and state unfair and deceptive acts or practices, also referred to as UDAP laws. The hackers themselves usually remain unidentified and escape civil liability.

Agentic AI attacks upend that model. The attacker is no longer an anonymous criminal group or individual, but an AI agent developed and trained by an identifiable company such as Anthropic, OpenAI, or Google. Logic would dictate that the attacker (and its developers) should be held liable for its part in the data breach.

Data-breach statutes haven't yet imposed liability on a separate company whose agent caused the breach. But agentic AI developers may still face liability under [Section 5](#) of the Federal Trade Commission Act — and, by extension, state UDAP statutes.

Defendants have argued that a “responsibility gap” exists because autonomous systems make decisions even their creators cannot reliably predict, making liability difficult to assign. On September 25, the FTC chair rejected that view, dismissing the idea that an AI agent is an independent actor that can break loose and absorb legal liability on behalf of developers.

The remarks signal that the FTC is prepared to bring enforcement actions against AI developers when their agents harm consumers, extending the agency’s existing Section 5 authority over data breaches.

State UDAP laws are largely modeled on the FTC Act and are often called “[little FTC Acts](#).” Courts applying state UDAP statutes are encouraged — and in some states required — to look to the FTC for guidance. That means agentic AI developers similarly could be held liable under state UDAP laws for cyberattacks carried out by their agents if they marketed the AI as safe or deployed it with inadequate safeguards despite known risks.

Maintaining Security Standards

For companies that fail to reasonably guard against agentic AI attacks, liability under data-breach and state UDAP statutes will likely remain unchanged.

First, data-breach statutes — for example, the CCPA — apply when specified personal information is accessed or exfiltrated because a business failed to maintain reasonable security. Because these statutes focus on the business’s security failures rather than the attacker’s identity, it can still provide redress after an agentic AI hack.

An AI agent generally can succeed only by exploiting an existing vulnerability the business failed to remediate. Now that these attacks are publicly known, businesses are on notice — which may raise the bar for what counts as reasonable security.

Finally, companies remain exposed under state UDAP laws based on their explicit or implicit statements and omissions about the strength of their cybersecurity or their compliance with state and federal law — even laws that lack a private right of action. When a company falls short of those representations, it may be engaging in unfair or deceptive practices, particularly where consumers relied on those representations when deciding to share their personal information.

This article does not necessarily reflect the opinion of Bloomberg Industry Group Inc., the publisher of Bloomberg Law, Bloomberg Tax, and Bloomberg Government, or its owners.

Author Information

[James Christie](#) is partner at Labaton Keller Sucharow whose practice focuses on complex securities fraud cases on behalf of institutional investors.

[Garrity A. Kuester](#) is an associate at Labaton Keller Sucharow whose practice focuses on cybersecurity and data privacy litigation.