

Anthropic AI model submitted false murder tip to Philadelphia police

Source copy kept by wheresthe.ai for the story "Anthropic AI model submitted false murder tip to Philadelphia police" (<https://wheresthe.ai/d/us-pa-anthropic-ai-model-submitted-false-murder-tip-to-phila-2026-10/>).

Source: <https://www.scmp.com/news/world/united-states-canada/article/3370414/anthropic-ai-model-sent-fake-murder-tip-us-police-hit-government-sites>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: 741d871509f26fd5bca54e579537767afea948c48e01851f0fab368ad89c75e

An artificial intelligence model developed by Anthropic submitted a fabricated tip about an unsolved murder to Philadelphia police, authorities said on Friday, criticising the company for taking two months to report the incident.

It is the first known instance in which a rogue AI appears to have tried to communicate a bogus tip to authorities, despite instructions not to create accounts or submit anything destructive, while not explicitly barred from submitting forms.

The Philadelphia Police Department said the false submission was made in July through PhillyUnsolvedMurders.com, a public website where people can share information about unsolved killings.

"I may have information regarding this case," Anthropic's model wrote in its submission. "I recall seeing someone matching the description in the area around (the street named on the page) during that time period. Please contact me if this information is relevant." The brackets featured in Anthropic's statement.

According to Anthropic's account, as relayed by police, the model was running a test that involved interacting with randomly selected websites when it reached the site and filed false information about an unsolved murder.

The AI model presented itself as someone who might have knowledge of the case.

Many of the cases Anthropic revealed involved websites run by federal, state and local agencies. Anthropic said it briefed the White House and notified all the agencies involved, but did not disclose who those parties were.

"Super intelligence companies must immediately disclose incidents involving their models and follow with swift, decisive action to remedy any and all harm," Federal Trade Commission director of public affairs Joe Gabriel Simonson said on social media.

This process was "not optional", he said, adding that the US government's newly created AI body, the Super Intelligence Force, would fulfil its responsibility.

The FTC said Anthropic disclosed to the SI Force on Friday its late September discovery of incidents involving what the task force called "unauthorised and fraudulent use of government and other systems".

The incident echoed other recent cases of unintended behaviour involving AI models, including one where an OpenAI agent undergoing a security evaluation broke out of its testing environment and breached systems at AI platform Hugging Face.

In September, Anthropic rival OpenAI apologised for the hacking of an Australian health data portal by a rogue AI agent, the first known instance of an AI agent exploiting a government website.

The episode heightened concerns about the AI industry's increased use of AI agents, systems programmed to take multi-step actions without human supervision.

Anthropic published a report on Friday outlining multiple types of "unintended" actions that its models have taken, including the incident involving the Philadelphia Police Department website.

Other organisations affected included the White House and other US government agencies, the report said.

People visit an Anthropic booth at a technology summit in San Francisco, California, in September. Photo: Reuters

The newly revealed incidents "had minimal real-world impact" and were "significantly less severe" than other cybersecurity incidents previously reported, Anthropic said.

The company outlined four categories of incidents that it found during an internal review of its Claude model: exploiting “basic” coding flaws, submitting forms on websites, bypassing requirements for tokens or fees and using short URLs to get around other limits.

In two cases described on Friday, Anthropic’s models obtained for free public data normally only available for a fee. Another instance revealed an obscure flaw that allowed use of a public tool hosted by a university.

Claude models were also able to bypass restrictions by using free services that shorten URLs.

Anthropic has turned off internet access for Claude during all internal testing for now “until we have confirmed that our security and monitoring measures ... reliably catch behaviours like these”, the report said.

Philadelphia police said the phoney tip, dated July 18, was flagged as spam and never reached the department’s Real-Time Crime Centre for vetting.

They added that there was no sign that police systems had been breached or department data compromised.

Anthropic discovered the incident on September 28, shut down the automated testing process responsible and added a new validation step for future tests, police said.

The company alerted the department on Wednesday, and the two sides met the following day.

“The two-month delay in detecting and reporting the incident to the City is unacceptable,” the department said.

Police said their safeguards had limited the impact, but that these “do not diminish the seriousness of an AI system presenting fabricated information as though it came from a person with knowledge of a homicide”.

“Unsolved cases involve real victims, grieving families and investigators working to secure answers,” the statement added.

Knowingly giving false reports to law enforcement authorities is a misdemeanour under Pennsylvania law, which specifies “a person”, however.

This includes “information relating to an offence or incident when he knows he has no information relating to such offence or incident”.

Additional reporting by Reuters