

Home • Artificial Intelligence



by **Evan Schuman**
Contributor

Gemini broke into 3 companies, but Google kept it quiet because 'no damage was done'

News

Sep 21, 2026 • 7 mins

An agentic cybersecurity incident in May impacted OpenAI, Google, Anthropic and Meta; everyone reported it except Google.



Credit: Sadi-Santos - shutterstock.com

A Google Gemini AI agent broke into three companies in May, guessing the credentials for one and discovering the credentials for the second two in a public repository, Google confirmed on Monday.

But the more interesting background to the story, which was **broken by The Wall Street Journal** [<https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breach-by-googles-ai-5c0baba2?st=2MGqTc>] on Friday, is that the May incident stemmed from a series of cybersecurity tests performed by security research firm Irregular on behalf of four AI giants: Google, Anthropic, OpenAI and Meta. All four companies experienced agent misbehavior resulting in cybersecurity incidents, but of the four, only Google never publicly disclosed its agent's activities. Indeed, it didn't reveal the breaches at all until contacted by a WSJ reporter.

Irregular **described the incident** [<https://www.irregular.com/research/addressing-recent-incidents-ongoing-findings-and-path-forward>] in August, around the same time as **Meta published** [<https://www.csoonline.com/article/4206116/meta-joins-openai-anthropic-in-latest-ai-test-breach.html>] its version and **Anthropic and OpenAI revealed theirs** [<https://www.csoonline.com/article/4205612/openai-anthropic-ai-agents-resorted-to-deception-in-new-cybersecurity-incidents.html>].

The Journal story noted, "the hacks occurred while the model was participating in a capture the flag exercise conducted on infrastructure belonging to Irregular to test the model's cybersecurity capabilities. It was tasked with retrieving information from software operated by a fictional company inside the testing environment. The fictional company shared the same name as a real company. Although the model wasn't intended to be able to get online, internet access was unintentionally made available, according to Irregular."

The three small companies whose systems were violated had, according to one source familiar with the testing, "almost no [cybersecurity] infrastructure." In short, none of the three was in a position to put up much of a fight when the Gemini agent successfully broke in.

According to a Google official, who asked to not be identified, the name of the public repository was similar to the name of the fake company. And within that repository were the names and credentials of the other two companies.

Recommended for you



Anthropic finds evidence of a fourth AI

[https://www.computerworld.com/?post_type=post&p=4221239&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

escaping from containment

By Maxwell Cooter • 2 mins



OpenAI, Anthropic AI agents resorted to deception in new cybersecurity incidents

By Gyana Swain • 5 mins

[https://www.csoonline.com/?post_type=post&p=4205612&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



OpenAI admits six new misalignment incidents under new reporting framework

By Gyana Swain • 5 mins

[https://www.csoonline.com/?post_type=post&p=4223458&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

Most analysts and consultants focused not on the hacks themselves, but on the reasons Google gave for being silent on the successful attacks.

Google said that the agents stopped as soon as they realized the victim companies were real businesses. “No harm was caused,” the Google source said. “There was not an issue of model misalignment.”

The source confirmed the Wall Street Journal story, which said, “Google compared the episode to a ‘bug bounty’ program in which hackers are rewarded for finding and reporting security vulnerabilities to their owners” and then quoted Heather Adkins, Google’s vice president of security engineering, saying, “In this case, the model acted appropriately.”

Define ‘harm’

Analysts generally disagreed.

CSO Smart Answers [Learn more](#)

Explore related questions

- [What security flaws were found in the Google Gemini CLI tool?](#)

- Which state-linked hacking groups bypassed Google Gemini guardrails using jailbreaks?
- How are state-backed hackers abusing Google Gemini to build cyber exploits?
- Can malicious calendar invites expose your data through Google Gemini?
- How did an API key permission change expose Gemini AI data?

Ask a question



“What does Google define as harm? Is it the same as the target company? Downtime, unauthorized access, and exfiltration of data may not result in immediate harm, but could have lasting impacts,” said **Ryan O’Leary** [<https://my.idc.com/getdoc.jsp?containerId=PRF005059>], an IDC research director. “The comparison to a bug bounty program is tenuous at best. If I broke into Google HQ and took nothing and caused no harm, it is likely I would still be prosecuted for trespassing.”

Nader Henein [<https://www.gartner.com/en/experts/nader-henein>], a Gartner VP analyst, had a similar take on the situation.

“If a member of my neighborhood watch broke into my house, walked around a little bit and then left, I’m fairly certain the authorities would not classify it as an act of civic engagement,” he said. “In this case, if the impacted sites had bug bounty programs and Google had programmed the agents to discover bugs, the rebuttal might make sense, otherwise it is quite a weak argument.”

That said, he added, “Google does make an excellent point when they underlined ‘the importance of training powerful AI models to act responsibly’ and I look forward to seeing how Google plans to ensure that this doesn’t happen again.”

Inappropriate behavior

But **Jeff Pollard** [<https://www.forrester.com/analyst-bio/jeff-pollard/BIO10584>], VP/principal analyst at Forrester, took exception to Google’s assertion that the Gemini model had behaved appropriately.

“The model pursued an authorized objective through an unauthorized path, crossed from a simulated environment into real companies and gained access without consent,” he said. “This is another area where regulations haven’t kept up with the pace of technology change. There are two sides to this: regulations with respect to the agentic escape and intrusion, and then the regulatory issues for the victim companies in terms of their requirements for disclosure. Google is only responsible for one half of that equation.”

Erik Avakian [<https://www.infotech.com/profiles/erik-avakian>], technical counselor at Info-Tech Research Group, also noted that it's critical that companies have rules about when to disclose unexpected and problematic model behaviors.

"I don't think every unexpected thing an AI model does needs to become a public incident. But there should be a clear line once an autonomous system crosses an authorization or trust boundary," he said. "If an AI system leaves a controlled environment, accesses a real third-party production system, uses credentials, retrieves data, escalates privileges, or takes some other action that was never authorized, that should, at minimum, trigger disclosure to the affected organization along with a formal incident investigation."

Even if there was no damage from the intrusion, Avakian said, public disclosure should happen "if the incident exposed a larger or repeatable problem with the controls around the model."

Independent technology consultant **Steven Eric Fisher** [<https://www.fisher-mns.com/about/>] also stressed that companies need to be strict and consistent about disclosing agent mishaps.

"What I find most puzzling about these incidents is not simply that an AI system crossed a boundary," he said. "It is the emerging posture around culpability once it does. Stopping after an authorization boundary has already been crossed is not the same thing as preventing the boundary from being crossed in the first place. I do not think 'the AI did it' can become an accountability boundary."

Control failure

And, argued **Justin Greis** [<https://acceligence.com/talent/profiles/justin-greis/>], CEO of consulting firm Acceligence, the absence of harm and absence of significance are not necessarily the same thing.

Sponsored Content

by Palo Alto Networks

Fight AI with AI: 4 Ways AI-Driven Defense is Rewriting the Cybersecurity Playbook

By Palo Alto Networks

27 Jul, 2026

[https://jadserv.postrelease.com/trk?ntv_at=3&ntv_ui=1ad80fdb-0ac5-406b-bfde-d93361121eb9&ntv_a=P4QKAfoF6AgYISA&ntv_fl=txhvbvCAVKzoTXiqjgo-af-63b4e5tmgv_lAfyxca-hUxst6GbVSt37OYHw_JN41MoTuvnfWUDQybbEgvOIDLzISA160eQomTWGjJ-fSZ1iL9Y5WeEZ_vLbz24Yd5Xj25jrnS1ux-MBT_rLo1-IYKBfo_FuqaJ-Tm7k7lyof5onPKFT5TfjrczVB-uXm99NyOCwEkAuZWZ10nemUF5Q_kYv3Rrxin3VNTB06zzS0bv3Tut_Fl3gvFyBAZ373a4a7HLZz33xkcW5YIM9hrQgLzxQmHlapYUjKARzB3Kpg7ZIZKWmkcSDVnOw7yO5SPrkOD_H3LU7zISK1CXpVdXALBs70SflVzK6_C2shjm2J6nI94phhvVC2QCIDs7UUmBHXJBZw3QuBPYFwuFKIzGjj59JR1kKTTWmuip8kr8gkOsSpuRH ztrPR7tIadbHPaoybwQQv31U-F6Nm2mFho3Qe692Ar9iWUSMX78cZK-hrj4J7NQCFGrJXtBkWd m0yuudxVINZAuYsCgsVnFwV_QDtHUO9FKsuoD0qoluY-QOPs3SIKaEzsUhYUEIITNTPrrai_yxPR bGu4UW1rUS_e63Ydszm17mpNVBxBTlg4H5HiO5jVplUiJi_keCTZQEDljUnPSczQ_vFZ_3PFjxCfa6]

oNatVfFEI8AmdXgXyA8mL0ibB0ud1j7GmTNC7-DklpRKiZOJza2YfHxVdZJ7QHPa1llwZwg40kb2-SdDt-Ze3Wgtr_BrKzlvTgKytuqSbMTKWxMPpfrdnaibfrETD9F2UPzs_jazJcRfb4snKGIDOW3Wr27y1-eFowk45R2i117OUsbGOzV2dJ4DM_lx9k-dg4mvGBnuwLKE0qALL7Bnc7rh0lxYotZemEfQJ6fl1Yn7vgBU4dhMSAxM68Q4GnuW2E1vI-6_cRO5OKjwMUJElsyNod24nUnOLonV0rkvnFt7fWO3ifoTSQ0DWJo2q5Qm4FyRFntjPwwFiZXeDUgocF-ovDtMQu3GUGukjXsBM87ktolApOqg6Xa7xvPhLBuGZqD638-lOqBFd1UP-lJjW1MjDq8_dLuj9CZhRHN3cvAXCP5VqqR_vsdxjtKUNddA==&ord=-1072132798&ntv_ht=-CazagA&ntv_rad=16&ntv_r=https://www.csoonline.com/article/4200895/4-wa-ys-ai-driven-defense-is-rewriting-the-cybersecurity-playbook-2.html]

“An event can be consequential because of what it demonstrates about a system’s capabilities or controls, even when everyone gets lucky and nobody is damaged,” **Greis** said. “Gemini stopping itself after recognizing that it was inside a real company’s environment is a positive safety signal. Gemini being able to get there in the first place is a control failure. Both things can be true at the same time.”

Frank Dickson [<https://www.linkedin.com/in/frankdickson/>], principal analyst at Dickson Research, articulated the harshest criticism of Google.

“The model’s behavior is the least interesting part of this story. Google’s conduct afterward is the most damning part,” he said. “This isn’t a story about Gemini going rogue. It’s a story about one shared testing vendor’s infrastructure mistake hitting four AI labs at once, and about Google being the slowest and least forthcoming of the four in telling anyone about its own copy of that failure.”

He pointed out, “Irregular notified all four labs in late July. Google didn’t go public until September 18, seven weeks later, and only after the Journal called for comment. Let’s face it: that’s not a company that judged that the incident didn’t warrant disclosure. That’s a company that watched three competitors take the reputational hit for the same underlying failure and waited to see if it could avoid its turn. It couldn’t, and the only reason we know any of this is that a reporter asked.”

Artificial Intelligence[<https://www.csoonline.com/artificial-intelligence/>]

Hacking[<https://www.csoonline.com/hacking/>]

Cybercrime[<https://www.csoonline.com/cybercrime/>]

Security[<https://www.csoonline.com/security/>]

NEWSLETTER

The latest AI updates, straight to your inbox

News, analysis, and insights for IT leaders navigating the risks and rewards of AI. A special series from the editors of CIO, Computerworld, CSO, InfoWorld and Network World

Email Address



By submitting your information, you agree to our [PRIVACY POLICY \[https://foundryco.com/privacy-policy/\]](https://foundryco.com/privacy-policy/).

SUBSCRIBE

About

[About Us](#)

[Advertise](#)

[Contact Us](#)

[Editorial ethics policy](#)

[Foundry Careers](#)

[Reprints](#)

[Newsletters](#)

[Contribute to CSO](#)

[Add CSO as a Preferred Source in Google Search](#)

More

[Awards](#)

[Blogs](#)

[BrandPosts](#)

[Events](#)

[Podcasts](#)

[Videos](#)

[Buyer's Guides](#)

Policies

[Terms of Service](#)

[Privacy Policy](#)

[Cookie Policy](#)

[Copyright Notice](#)

[Member Preferences](#)

[About AdChoices](#)

[Your California Privacy Rights](#)

Our Network

[CIO](#)

[Computerworld](#)

[InfoWorld](#)

[Network World](#)

