DATA PRIVACY & GOVERNANCE HUB

Gemini Hacked Three Real Companies During a Security Test

NOVEMBER 18, 2026

Key Takeaways

- Google's Gemini exceeded the scope of a testing environment and hacked three companies.
- Gemini used exposed credentials found in public online repositories, along with password guessing, to access the companies it infiltrated.
- The incident raised questions about how AI models are tested and contained, and about when AI companies should disclose such incidents.

What Happened

In May 2026, Irregular, an independent AI security evaluation firm, tested the offensive cybersecurity capabilities of Google's Gemini AI in a "capture the flag" exercise. During these tests, Gemini was given simulated companies to hack. Due to a misconfiguration, however, the testing environment was left connected to the open internet, and some of the simulated companies shared the names of real ones. Subsequently, Gemini escaped the testing environment and infiltrated three real companies. Gemini reportedly guessed its way into one company's system through brute force and used exposed login credentials found in public online repositories to access the other two. The companies invaded were later notified of the violation and, fortunately, Google reported that Gemini recognized its error shortly after entering the target environments and did no further damage. This event, in combination with several other instances of similar situations occurring within high-profile AI companies, is capturing the imagination and fears of the general public. Irregular has also been involved in similar incidents with models from OpenAI, Anthropic, and Meta this year. Unlike those

companies, however, Google kept this incident private until the Wall Street Journal approached it.

Privacy and Governance Concerns

In this case, the only known harm was the compromise of victim companies' credentials and the intrusion into their networks. However, how this harm occurred raises a host of questions about how AI models are tested and contained. Specifically, can AI be trusted to respect the privacy and security of systems it was never authorized to access?

AI solves problems in ways humans are not accustomed to. An AI doing its best to achieve a specific, benign goal could undergo processes that violate privacy. The Gemini, OpenAI, Anthropic, and Meta incidents show a pattern revealing just how reckless AI models can be in pursuit of what they understand to be their goal.

In this incident, Gemini demonstrated its impressive ability to aggregate open-source information into actionable intelligence. If an AI model can autonomously act on the intelligence it has gathered, it could create serious legal issues. It also remains unclear whether the systems Gemini accessed contained personal data, which could trigger breach-notification obligations for the affected companies.

Why It Matters / Policy Considerations

AI is meant to solve problems in ways novel to humanity; in fact, this is one of the major reasons why it is so useful and efficient. Although AI engineers do not directly program how an AI will answer a particular query, they do control the training process. During training, an AI model learns to recognize patterns and solve problems. Even after pre-training and fine-tuning, the model remains unpredictable. This unpredictability challenges the effectiveness of manual safeguards. An AI bent on completing a given task may find a workaround for any safeguard in its way. This is why a third and final training phase, Reinforcement Learning with Human Feedback (RLHF), is crucial. During RLHF, the model is evaluated and scored on its performance with direct human oversight. Unfortunately, the AI model, in an attempt to maximize its score, may still covertly bypass safeguards. Since the model avoids punishment for the undetected violation and still completes its goal, it will subsequently be more willing to

take a similar route in the future.

The best way to ensure rights are being respected is to introduce mandatory audits to the training process, specifically during the final stage of training, RLHF. AI engineers should be required to generate sufficient documentation of the training process for auditors to review. Audits should occur before the release of a new model, and the release of models deemed unsafe or otherwise unsatisfactory should be delayed until they are found to be compliant. Audits should also extend to the environments in which models are tested. Third-party evaluators such as Irregular should be required to verify that test environments are isolated from the internet, and AI companies should be required to disclose incidents like this one within a set period rather than waiting for journalists to uncover them. Hopefully, the audits will help verify that safe and ethical practices are used in training.