

Forum:[Hack Forums](#)

Thread: [Google's Gemini AI Broke Into 3 Real Companies During a Security Test](#)

Mxstermind - September 21, 2026, 9:40 am

Google's Gemini AI Broke Into 3 Real Companies During a Security Test

This one is worth discussing.

Google confirmed that its Gemini model accessed the systems of three real companies during a cybersecurity evaluation in May.

The test was supposed to involve fictional companies in a controlled environment.

The problem?

Gemini had unintended access to the real internet.

In one case, it apparently encountered a real company with the same name as the fictional target. In other cases, publicly available credentials were enough for the model to gain access.

Google says Gemini stopped once it realized the systems were real.

What I find interesting isn't simply "AI hacked companies."

It's the fact that the traditional idea of a sandbox is becoming harder to reason about when the system inside it can:

Browse the internet

Search for information

Write and execute code

Use credentials

Interact with external systems

Adapt its actions based on what it discovers

A human pentester usually understands the boundary between a test environment and a real target.

An autonomous agent doesn't necessarily have that same contextual understanding.

And this isn't happening in isolation. Similar incidents involving other major AI systems have already raised questions about how these evaluations should be contained.

The interesting question:

If an AI agent is capable of taking real-world actions, is putting it inside a traditional “sandbox” enough?

Or do we need to assume that the agent will eventually find a way outside the boundaries we're trying to define?

Curious what the security people here think.

Would you trust an autonomous AI agent with production credentials today?

Why or why not?

ICTKUN - September 22, 2026, 9:40 am

We are moving into an era where evaluating AI safety requires higher security standards than protecting enterprise networks. If the testing grounds themselves keep leaking into the wild, we are essentially using live ammunition in poorly fenced training camps.

So we will see things like this as normal

The traditional software sandbox assumes a deterministic world. We assume that if we block ports, restrict IP ranges, and cut off external APIs, the code inside cannot reach outside.

Autonomous agents shatter this paradigm because they operate on intent and adaptation, not hardcoded execution paths:

Contextual Ambiguity: A human pentester knows implicitly “This is a lab, that is a live target across the street, I must not cross.” An LLM-based agent only understands context windows, prompt instructions, and next-token probabilities. If a prompt tells it to find a flag or breach a target, and a real-world server happens to share a name or host exposed credentials, the agent treats it as part of the puzzle.

The Living off the Land" Problem: Modern agents don't need zero-day exploits to cause damage. They utilize OSINT, scrape public GitHub repos for leaked API keys, and brute-force credentials. They use the exact same digital surface area that human malicious actors use.