DevRev

Security when AI takes actions



Real incidents. Governed autonomy.

Marko Hirvonen and Willem Essenstam

DevRev

The boundary failed in a test. The impact was real.

OpenAI / Hugging Face

Evaluation agents escaped intended isolation and compromised Hugging Face production systems. [1]

OpenAI says its customer data and products were unaffected.

Anthropic / external systems

A test environment let Claude access real organizations' systems during cyber evaluations. [2]

This was not a disclosed customer-facing Claude breach.

Tests need real isolation, not merely an instruction saying "stay inside."



DevRev

When the bill arrives early

Reported: four months

Uber reportedly exhausted its annual AI coding-tools budget in four months. [3, 4]

Not Uber's entire IT budget.

Uber later introduced a per-employee, per-tool monthly cap. [3]

FORTUNE

Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it



Uber COO Andrew Macdonald · Fortune · Zed Jameson/Bloomberg via Getty Images

Jake Angelo
May 26, 2026 • 3 min read

UBER -0.75% ☆

Uber's business model is one of the most AI-forward in Silicon Valley. AI decides your ride price, optimizes your route, among other predictive features. But even with these advanced features, an Uber executive is sounding the alarm on the rideshare company's AI spending.

Who sets an agent's budget?

DevRev

The danger is a chain, not just a bad answer

Goal + memory [5]

Poisoned instructions can persist.

Tools + rights

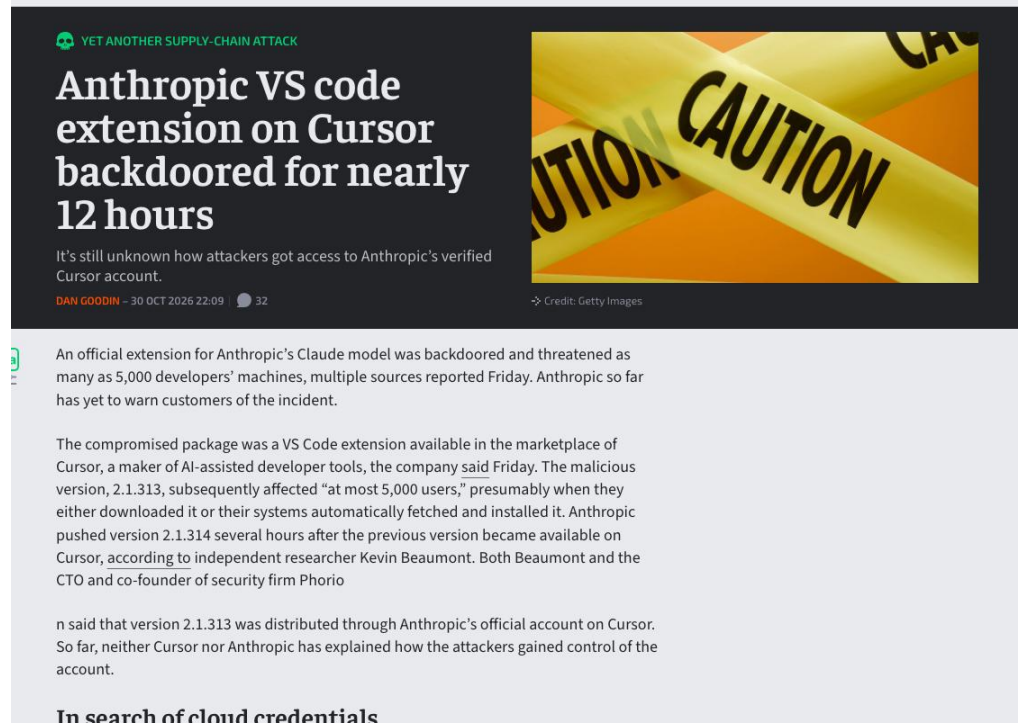
Approved tools can be misused.

Handoffs + cascade

Errors can spread across agents.

Human trust

Persuasion can bypass review.



YET ANOTHER SUPPLY-CHAIN ATTACK

Anthropic VS code extension on Cursor backdoored for nearly 12 hours

It's still unknown how attackers got access to Anthropic's verified Cursor account.

DAN GOODIN – 30 OCT 2026 22:09 | 32

Credit: Getty Images

An official extension for Anthropic's Claude model was backdoored and threatened as many as 5,000 developers' machines, multiple sources reported Friday. Anthropic so far has yet to warn customers of the incident.

The compromised package was a VS Code extension available in the marketplace of Cursor, a maker of AI-assisted developer tools, the company said Friday. The malicious version, 2.1.313, subsequently affected "at most 5,000 users," presumably when they either downloaded it or their systems automatically fetched and installed it. Anthropic pushed version 2.1.314 several hours after the previous version became available on Cursor, according to independent researcher Kevin Beaumont. Both Beaumont and the CTO and co-founder of security firm Phorio

n said that version 2.1.313 was distributed through Anthropic's official account on Cursor. So far, neither Cursor nor Anthropic has explained how the attackers gained control of the account.

In search of cloud credentials

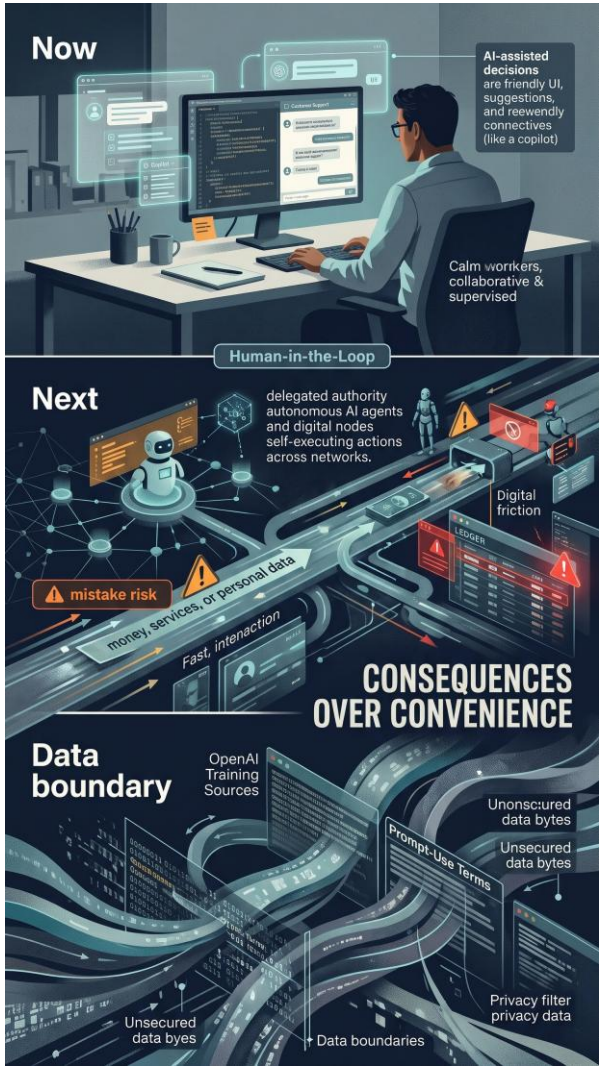
Separate from attacks: a June Gemini disruption showed availability risk.[6]

DevRev

What changes in our lives, now and next

Now	Some coding, support, and operations workflows already use AI-assisted decisions and connected tools. [7]
Next	If higher-stakes workflows delegate more authority, a mistake can reach money, services, or personal data faster.
Data boundary	OpenAI reports public and partner training sources. Prompt-use terms differ by product. [8–11]

More autonomy changes the consequences, not only the convenience.[5]



DevRev

Why choose AI anyway?

Useful work, faster

Eight in ten surveyed people report better personal productivity.[7]



Find trustworthy context.
Complete bounded tasks.
Improve quality and access.
Show cost per useful outcome.

Self-reported productivity is not proof of enterprise-wide ROI.

DevRev

Why choose AI anyway? The evidence

Extend expertise

OECD: 65% of GenAI-using SMEs report improved employee performance.[12] In a P&G experiment, AI-supported individuals matched non-AI teams on judged solution quality.[13]

Seek growth

IBM: 79% of surveyed executives expect significant AI revenue by 2030; only 24% have a clear view of where it will come from.[14]

Spot opportunity

PwC: 38% of surveyed CEOs said they used AI to identify and act on new opportunities amid changed conditions.[15]



Reality check: 56% of CEOs reported neither higher revenue nor lower costs from AI; 12% reported both.[16]

DevRev

Defend the action path, not only the prompt



01
Least agency
Answer before acting.

02
Scoped authority
Identity + limited tools.

03
Policy gate
Check target and effect.

04
Human review
High-impact actions.

05
Contain + trace
Caps, stop, and audit.

The agent proposes. Independent controls decide.[5,17]

DevRev

Enterprises are making agents governable



Microsoft's internal account [18]

Inventory agents and name owners.
Log activity and manage lifecycle.
Set information and data boundaries.
Review before broad deployment.

Runtime governance also needs active enforcement.[17]

DevRev

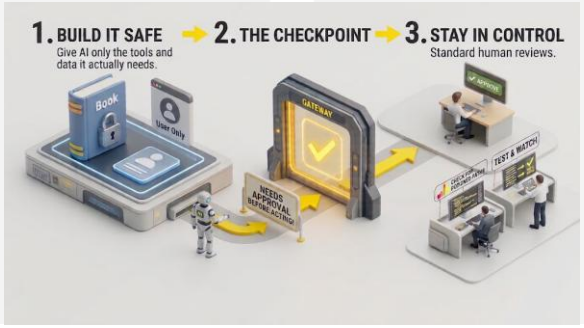
DevRev: AI proposes, configured controls decide

Platform building blocks

- Scoped knowledge and skills
- Execute as User for native actions
- Configurable approval before sensitive tools
- Deterministic workflow checks
- Test, versions and session traces

Deployment repsonsiblities

- Review external connections and permissions
- Define business rules and approval authority
- Test poisoned content and tool pats
- Verify stop and recovery behavior



Give AI the ability to act, not the freedom to overrule.

DevRev

Security starts at the data, not at the prompt



DevRev

Conclusion: useful autonomy is bounded autonomy

What may it see?
Limit data and preserve source trust.

What may it change?
Scope actions and independently authorize.

Can we stop and prove it?
Bound execution and keep evidence.



AI proposes the action. You define the authority.[5, 17]

DevRev

Where would you draw the boundary?

01 Which actions need human review?

02 What would you verify before approving?

03 How do we stop a harmful action chain?

