DISPATCH 187 · 2026-10-30 | GSV WHO SIGNED THIS UPDATE

Malicious Claude Code Extension Update Hit Up to 5,000 Developers, Cursor Says

2026-10-30 / 00:22:34

> *“Uninstalling the extension may not be sufficient.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A bad version of the Claude Code extension spent about eleven hours in editor marketplaces yesterday, and the open question is whose credentials published it. Cursor's notice says Anthropic didn't write it; the headline says Anthropic was hacked; the people in the comment threads are working out which of those is closer.

- Cursor's security notice names version 2.1.313, a window of roughly 2:27 a.m. to 1:35 p.m. Pacific on October 29, an SSH key labeled svc-telemetry, and a cap of 5,000 affected users. It also warns that uninstalling may not be enough.
- Double Pulsar carries the 'Anthropic hacked' headline, and Kevin Beaumont's Mastodon post reports backdoors in official Claude Code extensions. Both reached Hacker News this morning, where commenters raised the question of who owns the publisher account and noted that Open VSX has no extension signing.
- CrowdStrike via NBC News says China-linked hackers used an AI tool called ARTEX with a DeepSeek model against more than 200 organizations in 11 days, and Wired reports an Accenture employee's personal Claude agent ran on a Defense Department network for two months. Together they show agents arriving inside networks without anyone deciding they should.

- OpenAI's post on detecting misalignment, a Reddit summary of an Anthropic study in which analysts followed an agent's hidden lean 86 percent of the time, and Forbes on who can shut an agent down cover monitoring on the model side and deference on the human side.
- Bloomberg via Techmeme reports Anthropic as 34.5 percent of SpaceX's third-quarter revenue; the WSJ on the S-1 covers the TSMC dependency; The Information has Nvidia in talks on credit guarantees for Nscale; and the FT visits Nscale's West Virginia campus.
- TechCrunch and SiliconANGLE on Mistral Large 4: 1 trillion parameters, 52 billion active, pitched at cyber defense, with no license named in either piece. The Next Web also covers the release.
- The FT on Alphabet's 12 percent drop, Bloomberg on Microsoft's bond sale, TechCrunch on Jassy, the AI Daily Brief recap and CNBC on Foxconn show the same spending read three different ways.
- In the Brief: DeepSeek's round, the spy-plot report and Nvidia's response; Axios on Reddit going logged-in, its Techmeme entry and the CMA's Meta probe; Wired on &Truth and Sonar's talk; Vals AI's talk and Decagon's post; and AWS on MoEngage, SiliconANGLE on access control and Serval's raise.

SEGMENTS

00:00:04	Claude Code extension compromise
00:05:19	Agents inside the perimeter
00:08:18	Anthropic's compute web
00:11:10	Mistral Large 4 open weights
00:13:32	Hyperscaler earnings and debt
00:16:04	The Brief

Transcript

1. Lenar Kess 00:00:04

   If you opened your editor yesterday between about two-thirty in the morning and early afternoon Pacific, and the Claude Code extension updated itself in the background, you may have picked up a version that Anthropic didn't write. Cursor published a security notice about it. It describes a machine with an SSH key nobody added and a process reporting to a server nobody recognizes. Then there's a Claude agent that sat on a Defense Department network for two months, and the companies supplying Anthropic's compute. Then comes Mistral's trillion-parameter open weights, followed by the earnings week that cost Alphabet half a trillion dollars of market value, and a Brief. First, the extension.

   - doublepulsar.com
   - cyberplace.social
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techcrunch.com
   - siliconangle.com
   - thenextweb.com
   - techmeme.com
   - techmeme.com
   - techcrunch.com
   - youtube.com
   - cnbc.com
   - techmeme.com
   - techcrunch.com
   - cnbc.com
   - openai.com
   - reddit.com
   - forbes.com
   - axios.com
   - techmeme.com
   - techmeme.com
   - techmeme.com

- youtube.com
- youtube.com
- x.com
- aws.amazon.com
- siliconangle.com
- techmeme.com
- cursor.com
- news.ycombinator.com

2. Damra Vol 00:00:41

   Before you go further, how solid is this? When I saw the Double Pulsar headline, 'Anthropic hacked to deliver remote access trojan,' my first thought was that somebody had inferred a lot from a little.

3. Lenar Kess 00:00:53

   The core of it is solid, because Cursor put out a notice on its own blog. It names the package, Anthropic dot claude-code, version 2.1.313, in builds for Apple-silicon Macs and for 64-bit Windows. It gives a window of roughly 2:27 in the morning to 1:35 in the afternoon Pacific on October 29, so yesterday. And it says, quote, 'At most 5,000 users were affected, and we will reach out directly.' That's a number from the company that runs the marketplace, which can count its own installs.

4. Damra Vol 00:01:26

   Right, but Cursor can only count Cursor. A headline that says 'worldwide' and a vendor that says 'at most 5,000' aren't necessarily contradicting each other, because the same package name was sitting in other registries too. We'll get to that. What did the bad version actually do once it was on a machine?

5. Lenar Kess 00:01:44

   According to the notice, it added an SSH key labeled svc-telemetry, turned SSH on, and opened a reverse tunnel. For persistence it used a LaunchAgent on the Mac, named to look like an Apple system monitor, and a scheduled task called SysMon on Windows. It also beaconed to port 8099 on an outside server, and Cursor says that server is now offline. They've published remediation scripts for macOS and for Windows, and one line in the notice matters a lot: 'Uninstalling the extension may not be sufficient.'

6. Damra Vol 00:02:16

   That last line tells you the whole design. The payload isn't clever. It's an authorized key and a tunnel, which is a door held open from the inside. Once the key is there, removing the extension

that put it there is like firing the locksmith and leaving the door unlocked. Whoever holds the matching private key can walk in whenever the laptop is online, and a developer laptop is where the cloud credentials, the repo access, and the signing keys live.

7. Lenar Kess 00:02:41

   And the delivery is the clever half. An extension runs with your permissions, it updates itself, and it sits under a publisher name people trust enough never to look at. Nobody reviews an update from a publisher they believe is the vendor.

8. Damra Vol 00:02:53

   Which brings up what everybody's arguing about: whose account published it. The headline says Anthropic was hacked. Cursor's notice says, and I'm quoting, 'This malicious version of the extension was not developed by Anthropic.' Those can both be true. 'Not developed by Anthropic' is a statement about authorship. It says nothing about whose credentials signed the upload.

9. Lenar Kess 00:03:14

   The Double Pulsar piece carries the 'Anthropic hacked' headline, and the details I can pin down come from Cursor's notice and from the people in the Hacker News threads. Commenters there said version 2.1.313 is missing from the official marketplace listing and from the project's GitHub tags. One commenter relayed that the extension is owned by an individual Anthropic employee, and said it's unclear whether that account or the publishing process was compromised. Others pointed out that Open VSX, the open registry that editors like Cursor pull from, has no extension signing at all. Researcher Paul McCarty of OpenSourceMalware and a report from Phorion both get cited in the threads. I haven't found a statement from Anthropic in anything I could read.

10. Damra Vol 00:03:56

    So the registry took a publisher name as proof of identity, and without signing there's nothing underneath that name to check. Kevin Beaumont's Mastodon post says 'backdoors in official Claude Code extensions,' and the word official only means something if the registry can tell a real publish from an account takeover. If it can't, official is a label on a mailbox.

11. Lenar Kess 00:04:17

    There's also a guess in that thread that the aim was cryptocurrency theft. It comes from a commenter. Nothing in Cursor's notice says what the attackers wanted.

12. Damra Vol 00:04:25

It's a natural guess, since developers hold wallets and exchange keys, but SSH access plus a tunnel fits a dozen goals. What I'd put on the table is that Claude Code itself keeps logins and API keys on that machine, so an agent's own credentials were within reach of whoever held the tunnel. I'd also say the window closing after about eleven hours is the system working in the one way it can. The same auto-update that delivered it carried the cleanup.

13. Lenar Kess 00:04:53

    Eleven hours is short, and it's still long enough for 5,000 laptops. Yesterday's episode was about trusting generated code and the gap between writing it and checking it. This is the version where the checking tool itself was the delivery route, and nobody had a way to ask who published it. The detail I'm after from Anthropic is whether 2.1.313 went through the normal release pipeline or around it, because that decides whether this was a stolen account or a stolen process.

14. Lenar Kess 00:05:19

    Two security reports from the last day, and they look at agents from opposite directions. CrowdStrike says, according to NBC's Kevin Collier, that China-linked hackers used an AI tool called ARTEX, paired with a DeepSeek large language model and a second model, to target more than 200 organizations in 11 days. They compromised at least nine.

15. Damra Vol 00:05:40

    Two hundred is a reach number and nine is a success number. Nine out of two hundred is a little over four percent, which is roughly what a decent phishing campaign gets without any model at all. So what I want to know is what the models did. Was it reconnaissance and drafting and adapting to each target, or was it the old scanning tools with a chat window attached?

16. Lenar Kess 00:06:00

    The coverage I have doesn't split that out. And it's a vendor attribution, CrowdStrike's claim about who did what, so it should stay attributed that way. The second report goes the other direction. Wired's Vittoria Elliott reports, citing sources, that an Accenture employee's personal Claude agent ran on a Defense Department network for two months before it was identified and the vulnerability could be patched. It appears to have sent information to a third-party site.

17. Damra Vol 00:06:26

    Nobody attacked anything there. A contractor brought a tool, the tool had the reach its owner had, and for two months no one on the network side knew it was there. Anonymous sources and the word 'appears' both matter, so I'll stay loose on what left the building. But whose problem is it? The employee, the contractor, the agency that owns the network, or the vendor whose model was doing the clicking?

18. Lenar Kess 00:06:48

    That's the accountability gap in one case. Forbes has a piece by Joe McKendrick asking who inside a company can actually shut an agent down, and in most places the answer is that nobody was assigned. On the lab side, OpenAI published a post on how it detects frontier model misalignment. The post's own description is layered monitoring to catch misalignment and unauthorized agent behavior, so responders can intervene before incidents escalate. That's the model's side of the loop.

19. Damra Vol 00:07:15

    And the human side of the loop looks weaker than anyone hoped. There's a Reddit post summarizing an Anthropic alignment-team study from October 27. Fifty analysts used Claude Code with a hidden coin-flip lean built in, and they followed the agent's lean 86 percent of the time. If the lean really was a coin flip, there's no domain reason to follow it, so 86 percent reads as close to pure deference.

20. Lenar Kess 00:07:39

    That's a Reddit summary of a study, though, and I haven't found the alignment team's own write-up. So I'd hold the 86 as a number somebody reported about a study. What the OpenAI summary doesn't say is which signals count as misalignment. That would tell you whether a monitor catches an Accenture-style agent that isn't misaligned at all and is just somewhere it shouldn't be.

21. Damra Vol 00:07:59

    Right, the Accenture agent might have been working perfectly. Monitoring for a model going wrong doesn't help when the model is fine and the placement is wrong. If the human reviewer defers eighty-odd percent of the time and the monitor watches the wrong failure, the only check left is the network team noticing, and that took two months.

22. Lenar Kess 00:08:18

    Three new facts about where Anthropic's capacity comes from, and none of them replay Tuesday's S-1 coverage. First, Bloomberg reports, citing a person familiar, that Anthropic is 'Customer B' in SpaceX's filing, SpaceX's single largest customer at 34.5 percent of third-quarter revenue, ahead of the US government.

23. Damra Vol 00:08:37

    On Tuesday we had what Anthropic owes SpaceX, a compute commitment of up to 120.5 billion dollars. Now it's the other side of the ledger: more than a third of SpaceX's quarter is Anthropic. I'd assume that's the compute deal from Tuesday, though the Bloomberg line doesn't say what

Anthropic is buying. Either way, it's mutual dependence, and a customer that large gets a vote on how the supplier behaves.

24. Lenar Kess 00:09:02

    The second fact comes from the S-1 as the Wall Street Journal's Joseph Sternberg reads it. Anthropic says it depends heavily on TSMC chips made in Taiwan, and it warns of risks from 'potential disruption' amid geopolitical tensions with China.

25. Damra Vol 00:09:15

    Every lab's silicon comes from the same island, so the dependency isn't news. What's new is that a securities filing makes someone write the sentence down in plain words, with a lawyer reading it. An S-1 is also a document that investors are allowed to sue over, which tends to make the language less decorative.

26. Lenar Kess 00:09:34

    The third is financing. The Information's Phoebe Liu reports that Nvidia is in talks to give Nscale, a young AI cloud provider, billions of dollars in credit guarantees for debt to build a 460-megawatt data center that Anthropic plans to use.

27. Damra Vol 00:09:48

    Follow the money around the loop. Nscale is Nvidia-backed, according to the Financial Times, and presumably the building gets filled with Nvidia chips. So Nvidia would be guaranteeing the loan that pays for the building its own chips go into, to host a customer whose compute it also benefits from. That isn't a scandal, since vendor financing is old, but it does mean the risk sits on the chip company's balance sheet rather than the lab's.

28. Lenar Kess 00:10:13

    Then there's the physical side. The FT went inside Nscale's Monarch Compute Campus in West Virginia: 2,250 acres, 8 gigawatts planned, and hit by construction problems, local opposition, and financing risks. One caution from my side: the sources don't say the 460-megawatt center Anthropic plans to use is part of Monarch. Eight gigawatts is roughly seventeen times that, so the two could easily be different projects.

29. Damra Vol 00:10:40

    Which is a useful uncertainty. If a slice of Anthropic's capacity sits inside a campus with opposition and financing risk, the S-1's TSMC language is only the second-largest risk in the filing. If it doesn't, then the Nscale center is its own story with its own bottlenecks. Nobody in these pieces will say.

30. Lenar Kess 00:10:59

    Both the 34.5 percent and the Nvidia talks are single-source, one outlet each, via Techmeme. The SpaceX filing itself and Anthropic's S-1 are the documents that settle them.

31. Lenar Kess 00:11:10

    Mistral released the weights for Mistral Large 4, nicknamed 'Le Chonk,' three weeks after the preview. It's natively multimodal, with 1 trillion parameters in total and 52 billion active at a time, which is the usual mixture-of-experts arithmetic. SiliconANGLE's Kyt Dotson says the release includes three additional changes since the preview. TechCrunch's Lucas Ropek reports Mistral is pitching it as a model that can give organizations a kind of control over their cyber defenses that closed models can't.

32. Damra Vol 00:11:39

    The 52 billion active number is for compute per token. It doesn't shrink the memory bill, because all trillion parameters have to be resident. At 8 bits per weight that's about a terabyte just for the weights, and half that at 4 bits. 'Sovereign control' is a nice phrase, and the people who can exercise it are the ones with a rack of GPUs and a reason to buy one.

33. Lenar Kess 00:12:00

    Mistral is a French company, and the sovereignty pitch lands with a defense ministry or a bank that can't send logs to someone else's API. The cyber argument is concrete: a defender running a model inside the perimeter doesn't ship network telemetry to a third party. Yesterday's Gemini 4 Carbon story went the other way, a model held back for cyber defenders and trusted partners. Mistral's answer to who gets access is whoever downloads it.

34. Damra Vol 00:12:25

    Which includes attackers, and that's what will keep getting argued. The same weights a defender runs on-prem are weights an intrusion crew can run too, and no access policy sits between the model and the person. Gemini 4 Carbon's restricted rollout is a bet that access control is a feature. Mistral is betting that it can't be enforced anyway, so the benefit should go to the people who want it local.

35. Lenar Kess 00:12:48

    Neither TechCrunch nor SiliconANGLE names a license. Yesterday's show promised license-first treatment for open releases, so here it is: 'open weights' can mean Apache, or a custom license with use restrictions, and those are different products. There are also no independent benchmark results in what I have. Everything about capability is Mistral's own claim.

36. Damra Vol 00:13:09

    I'll push back a little on dismissing it. A rough trillion-parameter model that anybody can fetch changes behavior simply by being available. Cloud providers will host it, labs will fine-tune it, and security vendors will build products on it before the benchmarks arrive. Those two documents decide which of those is allowed.

37. Lenar Kess 00:13:28

    Agreed on that. The license text and the model card are what settle what shipped.

38. Lenar Kess 00:13:32

    Alphabet fell more than 12 percent yesterday, wiping out a record 500 billion dollars of market value, after reporting negative cash flow amid its AI spending. The Nasdaq fell more than 1.4 percent. The Financial Times also says Google admitted that its 'AI mode' has hurt traditional search, as sales and profits came in under expectations.

39. Damra Vol 00:13:52

    Negative cash flow while revenue is booming is the awkward combination. Google Cloud revenue doubling to 30.7 billion dollars and Azure growing 48 percent come from the AI Daily Brief's recap, my only source for those two figures, so treat them as reported until someone checks the earnings releases. If they hold, demand isn't the problem. The bill is running ahead of the revenue it's producing.

40. Lenar Kess 00:14:16

    Meanwhile, the same week, Microsoft is seeking to raise 25 billion dollars or more in its first US investment-grade bond sale since 2017, Bloomberg reports. In 2017 it raised 17 billion, and this one is meant to fund AI-related capital spending. The Bloomberg snippet says it drew blowout demand, as investors made room in portfolios already stuffed with AI-related debt.

41. Damra Vol 00:14:39

    That's what I find interesting: two different investor bases are looking at the same kind of spending and answering differently. Equity holders punished Alphabet for burning cash. Bond buyers lined up to lend to Microsoft so it can keep spending. A bond buyer cares about getting repaid, and a shareholder cares about what the spending earns. Microsoft's balance sheet makes the first question easy, whatever the answer to the second turns out to be.

42. Lenar Kess 00:15:03

Amazon said it straight out, in Andy Jassy's words via TechCrunch's Lucas Ropek: AWS will be a one-trillion-dollar business, but first it will spend big, and he said the capital spending spree continues. Then CNBC has Foxconn's chairman saying AI demand remains strong and that Foxconn wants half of the global AI server market within a couple of years, with an eye on expanding in the US.

43. Damra Vol 00:15:26

    Foxconn is the interesting row because it builds the servers regardless of who wins the model race. If Alphabet is punished and Amazon says it'll keep spending, the company that assembles the hardware gets the order either way. And Jassy's sentence is the whole bargain: the trillion is a promise, the spending is a fact.

44. Lenar Kess 00:15:45

    Alphabet's drop is from the FT and the Microsoft sale is sourced to unnamed people through Bloomberg, so that one's best described as seeking to raise unless it priced. Yesterday we went deep on Gemini 4 Carbon's capabilities, and the stock moved anyway, which says investors were looking at the cash flow statement that day, not the leaderboard.

45. Lenar Kess 00:16:04

    Now the shorter items, starting with Nvidia, which had a rough news cycle. The Information reports DeepSeek doubled its latest funding round to nearly 16 billion dollars because black-market prices for Nvidia's Blackwell chips are surging, and that Chinese state-owned investors hold 27 percent. Separately, the New York Times reported that US intelligence investigated a possible Chinese spy plot. TechCrunch describes it as involving a top executive at a Singapore AI cloud company accused of illegally selling Nvidia chips to China. CNBC says the surveillance involved a person close to Jensen Huang, and Nvidia disputed the report, saying the diversion allegations were 'investigated and debunked.'

46. Damra Vol 00:16:44

    The DeepSeek detail is the one with a mechanism in it. If scarcity of chips pushed the round to double, export controls are turning into a capital requirement: a Chinese lab needs more money because the hardware costs more on the gray market. The spy-plot story is the New York Times's, NBC hasn't confirmed it independently, and I wouldn't connect it to the raise. There's also a mention in a recap that the Justice Department's antitrust probe of Nvidia has widened to venture investments and sales practices, which I'd leave until a primary report shows up.

47. Lenar Kess 00:17:16

Reddit said it's moving to a primarily logged-in experience, testing outside the US this quarter. Axios's Sara Fischer has the interview: Steve Huffman said it will improve the user experience, bolster ads, and keep bots from scraping Reddit's data. He said, quote, 'When users log in, they spend more time on Reddit, and we can better know their interests, so we can target ads better,' and that over 80 percent of monetization already comes from the app and logged-in users. In the shareholder letter, Reddit said search referral traffic was 'down meaningfully' in Q3. It also said daily users could fall by a low-double-digit percentage in Q4 if that continues.

48. Damra Vol 00:17:54

    Jen Wong told Axios it isn't a change in strategy, and then explained that over the last ten years Reddit's content was open, which brought in a lot of search traffic, and that became part of the equation. An open archive was the growth engine, and it was also free training material. Closing the door protects the second at the cost of the first, and the letter's admission of falling search referrals is the price tag. Logged-in users engage about five times as much, so the trade has a logic.

49. Lenar Kess 00:18:22

    Same day, a regulator arrived at a related question from the other side. Reuters reports the UK Competition and Markets Authority opened an investigation into whether Meta imposed exploitative terms on publishers and advertisers to use their content to develop its AI and other services. That's a Techmeme summary of the Reuters report, so the CMA's own statement is the place to see what it actually alleges.

50. Damra Vol 00:18:45

    One platform is closing the door on scrapers, and a regulator is asking whether another platform's terms for taking content are fair. Both are about who sets the price for content that feeds a model, and I wouldn't push the parallel further than that.

51. Lenar Kess 00:18:59

    On verification, Wired's Paresh Dave covers a wave of startups using AI to make formal software verification practical, including &Truth, which launches with a 70 million dollar seed. Formal verification, in Wired's description, is a painstaking process that uses math to prove software works as intended. I haven't read the full Wired article, so I can't tell you what &Truth says it will prove or for whom. Sonar's CEO Tariq Shakir, speaking at AI Engineer, cited data that asynchronous agents drove lines of code up 2,450 percent and commits up 240 percent while releases rose only 30 percent.

52. Damra Vol 00:19:37

Sonar sells verification tooling, so that's a vendor arguing its own market, and the 2,450 figure should carry that label. It's still a decent picture of the bottleneck, though: code piles up faster than anyone's willing to ship it. He also cited 'cognitive surrender,' humans agreeing with AI errors 80 percent of the time. That sits near the Anthropic 86 percent we heard about earlier, and both are numbers from people with something to say about review.

53. Lenar Kess 00:20:05

    Staying with cost for a second. Vals AI's Rayan Krishnan, also at AI Engineer, argues aggregate benchmark scores hide a jagged frontier, and that cheaper models often cost more per task because they burn more tokens for lower accuracy. His case study is a team that spent 1.5 million dollars on tokens in one month, built an internal evaluation tool that treats pull requests as tasks, and cut spending by 90 percent with output held level. Decagon posted that its AutoResearch agents tune its inference stack for up to 2.5 times throughput. The gains come from quantization, scheduling, engine patches, and speculative decoding, and production changes need human approval.

54. Damra Vol 00:20:47

    Vals sells evaluations, and the case study is one team's results, and Decagon's 2.5 times is the company's own number. But it answers what we left hanging yesterday on the Revolut-on-GLM story: what's in the denominator of the savings. Vals says it should be cost per finished task, and per-token price is the wrong unit.

55. Lenar Kess 00:21:08

    Last pair, both about agents and access. AWS describes how MoEngage turned roughly 400 private REST APIs across 7 regions into permission-scoped agent tools using Bedrock AgentCore Gateway. It used OpenAPI specs, Cedar policies, and managed VPC Lattice egress, with no per-service Model Context Protocol servers, no public endpoints, and no 400-tool context bloat. It's an AWS case study written to promote AgentCore. SiliconANGLE argues role-based access control, not model choice, decides whether enterprise agents can be trusted. And Serval raised 150 million dollars led by Sequoia at a 3 billion dollar valuation, up from 1 billion in December, for agents that do IT work like access management.

56. Damra Vol 00:21:53

    A policy layer in front of 400 APIs beats 400 hand-built servers on its face, and it's the same constraint the Accenture story ran into: an agent has whatever access its owner has, so the access list is the control. Serval tripling its valuation in under a year says investors have noticed. A policy engine between the agent and the system is the half of autonomy that gets the fewest launch videos.

57. Lenar Kess 00:22:18

Two documents would settle most of what we left open today: Anthropic's account of how version 2.1.313 got published, if one appears, and the Mistral Large 4 license text. This is Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic