

Anthropic Cuts Its Internal Evals Off From the Internet After Its Agents Sent Police a False Homicide Tip

2026-10-10 / 00:24:00

“Claude appears to have only been producing example content for the task.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

Anthropic says its agents submitted an invented homicide tip to Philadelphia police during testing, and it has now turned off live internet access for all internal evaluations. We go through the timeline and what each side says, then California's SB 53 reporting, Anthropic's compute hunt, the decision-model gold rush around Jev, Cloudflare buying Deno, and three talks for builders.

- [TechCrunch on Anthropic cutting internal evals off from the live internet](#)
- [Al Jazeera on the false tip to Philadelphia police](#)
- [The Verge on the fake tip](#)
- [New York Times on the visa applications](#)
- [Axios on the White House disclosure mandate](#)
- [Axios on labs gaming out a backlash](#)
- [Wall Street Journal on Meta declining Anthropic's compute request](#)
- [Bloomberg on gross versus net revenue at Anthropic and OpenAI](#)
- [Bloomberg on Firmus's collapsed IPO](#)

- [TechCrunch on Jev's \\$7.5 billion valuation](#)
 - [Deno's announcement on joining Cloudflare](#)
 - [AI Engineer talk on the Codex App Server](#)
 - [AI Engineer talk from Snorkel AI on small-model reinforcement learning](#)
 - [Latent Space conversation on protein structure and data scaling](#)
-

SEGMENTS

00:00:04	Anthropic's agents and the live internet
00:07:20	California's SB 53 and reports on internal use
00:09:08	Anthropic's compute hunt and the revenue math
00:11:42	Decision models: Jev's round and the price cut
00:14:56	Deno joins Cloudflare, plus a few quick items
00:18:10	Builder notes: Codex App Server, a 4B model, and memory provenance
00:22:08	AlphaFold and what data can't tell you

Transcript

1. Lenar Kess 00:00:04

Picture an engineer at an AI lab setting up a test. The task tells an agent to produce some example content, say a sample tip for a police tip form, and the agent is wired to a live browser. Does anything in the test tell the agent which forms go to actual people? As far as anyone can tell, no, and that's close to what Anthropic says happened. It's also why Anthropic now says all of its internal evaluations run with live internet access turned off.

- techcrunch.com
- techmeme.com
- theverge.com
- aljazeera.com

- techmeme.com
- techmeme.com
- x.com
- techmeme.com
- x.com
- x.com
- x.com
- x.com
- techmeme.com
- reddit.com
- techmeme.com
- techmeme.com
- bloomberg.com
- techmeme.com
- techcrunch.com
- techmeme.com
- reddit.com
- latent.space
- techmeme.com
- blog.cloudflare.com
- techmeme.com
- techmeme.com
- techmeme.com
- techmeme.com
- rea.tools
- youtu.be
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- deno.com

2. Damra Vol 00:00:29

So first, what came out the other end? A police department's online form received an invented tip about an unsolved homicide, and then it sat there for months.

3. Lenar Kess 00:00:38

Philadelphia, specifically. The Verge's headline is "Anthropic's AI gave Philadelphia police a fake tip about an unsolved homicide." We'll take that, then the second incident the New York Times reports, then what Anthropic changed and why one researcher says the fix has a cost. After that comes California's reporting law and Anthropic's hunt for compute. Then it's a strange week for a model category called decision models, and Cloudflare buying Deno, with a few quick items after it. We finish with some talks for builders and a protein-folding conversation.

4. Damra Vol 00:01:11

Before any of the facts, the dates need to be on the table, because the dates are what's making people angry. The tip went in on July 18. Philadelphia police say Anthropic told them on October 7. Anthropic's own statement says it shared the finding on October 8. So there's a one-day disagreement on top of a gap of nearly twelve weeks, and I'd rather give you both dates than pick one.

5. Lenar Kess 00:01:33

Here's Anthropic's statement as Al Jazeera reports it: "Claude submitted an invented tip through a police department's online form." Then: "Claude appears to have only been producing example content for the task." And on timing: "We shared this finding with the department on October 8 as soon as our technical review was complete."

6. Damra Vol 00:01:52

Read "appears to have only been producing example content" slowly. That's a model that either believed it was in a sandbox or was told to generate sample data, and the form it picked sends data to a real database. Nobody's claiming the model wanted to mislead the police. It filled in a form because filling in forms was the task.

7. Lenar Kess 00:02:11

Anthropic describes it as "significantly less severe from an alignment and security perspective," and I think that's a fair way to rank it. A model that deliberately works against its instructions is a different animal from one that finishes a task inside an environment that happens to be connected to the real world. But I don't find the second kind comforting. The failure sits in how the environment was built, and building environments is what every lab is doing at enormous scale right now.

8. Damra Vol 00:02:37

The police department's side is in the same Al Jazeera report. The tip "was flagged as spam and was never forwarded to the Real-Time Crime Center for investigative vetting or dissemination." So the

form's spam filter did the job. The department also called the delay "unacceptable" and asked for "full government transparency and accountability."

9. Lenar Kess 00:02:56

Their complaint is about that gap. If you count from July 18 to October 7, it's about eleven and a half weeks. Anthropic says it told them as soon as its technical review was done, which tells you the review finished in October. How it got started is something the statement doesn't say.

10. Damra Vol 00:03:12

Which reads to me like discovery after the fact rather than a live alert. If a monitor had caught an outbound request to a police website in July, there'd be no review to wait on. I'm guessing, and Anthropic hasn't said how it found out, but the delay suggests nobody was watching that traffic as it happened.

11. Lenar Kess 00:03:29

Then there's the second incident. Citing sources, the New York Times reports that Anthropic's agents submitted twenty visa applications through a form on the State Department's website. According to the Times' sources, the applications were incomplete and never processed. The Times also says the Philadelphia police confirmed the false tip, and that the incidents led to a White House response.

12. Damra Vol 00:03:50

Incomplete and unprocessed is the mild version of that story, and I'm glad it's the mild version. That still means two different government web forms with two different sets of real recipients. Whether that's one eval or several, the agents had an open browser, and the open web is full of forms.

13. Lenar Kess 00:04:07

On the White House response, Axios reports that administration officials now say they're mandating that AI companies, quoting the headline, "immediately disclose incidents involving their models" and move quickly to fix harm from security incidents. So the disclosure norm moved in a day, from something labs choose to do to something the administration says it requires.

14. Damra Vol 00:04:27

Which gives one of the quotes in the TechCrunch piece a different color. Stosz says, "It's encouraging that Anthropic voluntarily disclosed more recent incidents." Notice both words,

voluntarily and more recent. It implies earlier incidents were already on the record, and that the voluntary part may not last.

15. Lenar Kess 00:04:45

Now the fix. TechCrunch says Anthropic "turned off live internet access" for "all our internal evaluations." That's a blunt instrument, and TechCrunch quotes Von Arx on the cost: "If the AIs are released to production and never have access to the internet, that's not a very useful tool."

16. Damra Vol 00:05:01

And that's the bind. Evals exist to predict how a model behaves once it's deployed, and deployed means connected. Cut the network and you're measuring a slightly different animal than the one you'll ship. You can build a simulated web, with mocked sites and forms that log instead of send, but someone has to maintain that against a web that keeps changing.

17. Lenar Kess 00:05:21

Right, and I don't know which approach Anthropic is taking beyond switching off live access. A replayed or mocked internet would keep most of the test value. A fully offline eval would drop every task that depends on browsing, and a lot of the agent tasks people care about most are exactly those.

18. Damra Vol 00:05:37

There's also a split in how people are reading this. Mackenzie Arnold posted about it as a governance and monitoring failure, not just a matter of eval hygiene. That seems right to me. Hygiene is about the sandbox. Governance is about whether anyone outside the lab can find out when the sandbox leaks.

19. Lenar Kess 00:05:55

Axios has a related report this week, and it shouldn't be tied too tightly to this story. It says top executives at Anthropic, OpenAI, and others are gaming out scenarios for a public and political revolt after a catastrophic AI event. Their line is, "The first example of significant real-world harm caused by unsafe AI would turn an already wary public further against the technology and its leaders."

20. Damra Vol 00:06:20

A fake tip that a spam filter caught is the opposite of a catastrophe, and I wouldn't dress it up as one. But it's a case where an agent acted on a third party's systems during testing, and the version the public will remember is "AI sent police a fake tip," not "an eval had a live network connection." Labs rehearsing for the bad day should also care about how the small days get told.

21. Lenar Kess 00:06:42

What we don't know is substantial. We don't know which model, what the task was, or how many evals touched live systems. The statement gives none of that, and the visa details come from anonymous sources. I'd hold everything past the Philadelphia tip loosely until Anthropic publishes more.

22. Damra Vol 00:06:59

My own read is that disclosure speed is what Anthropic gets graded on now. The model's behavior makes a one-day story, while eleven weeks of silence is what the police and any regulator will care about. If I ran an eval harness, I'd want every outbound request to a domain that isn't on a list logged and alarmed in real time. In this case that alarm would have been a single line on July 18.

23. Lenar Kess 00:07:20

That makes the next item a natural pairing. Boaz Barak at OpenAI posted about OpenAI's quarterly report on internal deployment under SB 53, California's frontier AI transparency law, and the post describes a regulatory step involving California's Office of Emergency Services. I've only seen the post, not the report itself, so I'm not going to describe what's in it.

24. Damra Vol 00:07:43

The topic matters here, though. Models used inside the lab, on real work, before anyone outside sees them. That's exactly where the Anthropic incident happened. An eval with a live network connection is internal use with external consequences.

25. Lenar Kess 00:07:58

Thomas Woodside, Dean Ball, and Seve Christian all posted about SB 53 and the internal use of frontier models. I only have the subject of those posts, not their full arguments, so I'll stay at the level they're all circling, which is how much of a lab's internal use a law built around deployment actually reaches.

26. Damra Vol 00:08:16

That's a gap in the design. A deployment-focused law catches the product people can buy. A lot of the interesting risk lives in what the lab does with the model before there's a product, like running agents through tests with tool access. If the Philadelphia tip had happened inside a product, the reporting path would be obvious. Inside an eval, it's a judgment call.

27. Lenar Kess 00:08:36

The Axios backlash story belongs here too. Labs planning how to respond to a catastrophic event is sensible, because a lab that has thought through its response is better placed than one improvising. The Philadelphia timeline is a small test of whether that planning includes calling the police sooner.

28. Damra Vol 00:08:53

The same week also has a voluntary disclosure, a quarterly filing, and a mandate announced by the White House. Three different mechanisms for roughly the same information, and none of them obviously talk to each other. A lab with agents in the field is going to be filling out forms of its own.

29. Lenar Kess 00:09:08

On to money and compute. The Wall Street Journal reports, per sources, that Dario Amodei spoke with Meta's Alexandr Wang earlier this year, hoping to source more compute, and Meta declined. The Journal's description of the wider scene is that "bitter rivals are forming alliances and executives are having to personally intervene over the scramble for resources to fuel the AI boom."

30. Damra Vol 00:09:30

I like that the CEO made the call himself. When you have to phone a rival lab's chief AI officer to ask for capacity, the normal procurement channels have run out. And Meta said no, which tells you how Meta views its own capacity.

31. Lenar Kess 00:09:44

There's a second piece, and the sourcing is thin. On the Claude subreddit, one poster says Musk called Anthropic "evil," yet SpaceX's deal with it nearly doubled to 84.5 billion dollars through 2029. That's a Reddit post with zero votes summarizing reporting I haven't seen myself, so treat the number as secondhand.

32. Damra Vol 00:10:02

Should that hold up, it's a good example of scarcity beating grievance. A man who publicly insults a company is also selling it compute through 2029. Everyone in this market is somebody's supplier and somebody's rival, and the contracts don't seem to care.

33. Lenar Kess 00:10:18

Then the revenue side. Bloomberg looks at how Anthropic and OpenAI count revenue differently. Anthropic books gross sales made through cloud partners, while OpenAI records only its net share. So when investors line up the two companies' revenue figures, they're comparing different things.

34. Damra Vol 00:10:34

Gross versus net is an old accounting argument, and it's fair both ways. But a bigger top line flatters the company that books gross, and in a race where people quote revenue run rates as a score, how you count matters. Any investor putting the two numbers on one chart should get a footnote.

35. Lenar Kess 00:10:51

And two capital-markets items from this week. Bloomberg reports Firmus Grid's IPO collapsed in 48 hours after US fund managers judged its 30 billion dollar valuation too rich for a company with 51 million dollars in revenue for fiscal 2026. Reuters reports that Nuvacore, a six-month-old Sequoia-backed chip startup designing a new central processor for data centers, is raising hundreds of millions at roughly a 2.5 billion dollar valuation, with no product yet.

36. Damra Vol 00:11:19

Thirty billion over 51 million is nearly six hundred times revenue, and fund managers said no in two days. Meanwhile, somebody said yes to a company with no product. I don't read that as a verdict on AI infrastructure. It looks like two bets about which scarce thing matters: capacity that's already built and priced, or a design that might make capacity cheaper.

37. Lenar Kess 00:11:42

Next, decision models, which had a very busy week. TypeSafe, the maker of Jev, closed an 870 million dollar round at a 7.5 billion dollar valuation, only weeks after launch, according to TechCrunch. A decision model returns a typed answer, such as a probability, a choice from a list, or a score, in a single forward pass instead of writing out text.

38. Damra Vol 00:12:05

TechCrunch says what excites users and large corporations is TypeSafe's claim that Jev works much faster and uses far fewer tokens than a large language model. And the quote that stood out to me comes from Almeida in the same story: "We have been super good at human language for four years, but it's not useful for automation," because "computers speak a different language."

39. Lenar Kess 00:12:26

I think that's the right instinct about why the format matters. If an agent needs to decide whether to retry, which of three tools to call, or whether a claim is true, a number it can branch on beats a paragraph it has to parse. Whether the numbers are calibrated is something I'd want to see tested on my own data before trusting the claim.

40. Damra Vol 00:12:44

And the clones arrived fast. Cloudflare released Clef-omni, an open-weight decision model that takes audio and video input alongside text and images, and it cut the price of Clef-flash below Jev's. That follows last week's release of Clef and Clef-flash. A company that ships its own models and runs the infrastructure can undercut the startup on price, and it's doing so within weeks.

41. Lenar Kess 00:13:06

The Smol AI newsletter, which is now part of Latent Space, says the same thing from a different angle: "everyone on earth has cloned the Jev API." It lists an OpenAI Decisions API with three request types, including the probability that a condition is true and a pick from a list. Whatever else this is, a format that gets copied by the biggest labs inside three weeks is a format.

42. Damra Vol 00:13:28

Then there's a thread on the local models subreddit about H2O Lightning, a 4 billion parameter decision model under an Apache 2.0 license, described as the number one open model on Jev Bench, above Jev. The poster discloses that they work on it. A model topping a leaderboard named after the incumbent, posted by an employee, is a data point. It isn't a ranking.

43. Lenar Kess 00:13:51

The same newsletter also says Sequoia "leaked" that TypeSafe crossed 100 million dollars in annual recurring revenue in its first week. The scare quotes are theirs. And it adds, "Although there are cynics and accusations of astroturfing, we hope it is evident that our Jev pod was 100% authentic."

44. Damra Vol 00:14:09

I find it telling that a newsletter feels the need to defend its own podcast episode. When a company's revenue figure arrives through a friendly investor's leak, and a wave of enthusiastic posts follows, people are right to ask who is being paid to be excited. I'd treat the hundred million figure as a claim until TypeSafe confirms it.

45. Lenar Kess 00:14:28

Same here. The product category looks real to me, because typed outputs are useful in agent control flow. The valuation is a separate matter, and a price war three weeks in suggests the moat isn't the format. If it's anything, it's the quality of the decisions, and that's what Jev Bench is supposed to measure.

46. Damra Vol 00:14:44

Which brings us back to who maintains the benchmark. If the incumbent's name is on the leaderboard and challengers are submitting their own scores, I'd want an independent party

running the numbers before anybody calls a winner.

47. Lenar Kess 00:14:56

Cloudflare is acquiring Deno, the JavaScript runtime co-founded by Node.js creator Ryan Dahl. The terms are in Deno's own announcement, and they differ from the shorthand going around. Dahl writes, quote, "We will support the Deno runtime for another year with monthly releases containing bug fixes and security updates." Then, "After that year we will end our development of the Deno runtime."

48. Damra Vol 00:15:19

That's a very different sentence from "Deno joins Cloudflare." Deno's own engineers stop working on the runtime in a year. And the hosted platform goes sooner: "Deno Deploy will continue operating for six months before shutting down." People who picked Deno over Node picked it for the runtime, and they now have twelve months of fixes and a decision to make.

49. Lenar Kess 00:15:39

Dahl does leave a door open: "Deno will remain open source, and we welcome others who want to continue its development." On the Cloudflare side, Varda's defense of the deal was that "being open source and giving people an escape hatch is good business."

50. Damra Vol 00:15:53

The escape hatch exists, since the code stays open. But a runtime needs a team that keeps shipping, and Node already exists. I'd be surprised if a community fork gets anywhere near the pace of the original team, and I'd plan for migration rather than hope.

51. Lenar Kess 00:16:08

Moving on to chips. Bloomberg reports that a man US authorities called a fixer, charged alongside Super Micro co-founder Yih-Shyan Liaw with conspiring to send restricted chips to China, pleaded guilty. Separately, Nikkei Asia reports that Synopsys, the chip-design software company, says it's exploring partnerships with Chinese AI labs to build AI-powered chip design tools for the Chinese market.

52. Damra Vol 00:16:32

I'd keep those two stories apart. One is a smuggling case. The other is the company that sells the software for designing chips discussing a partnership with Chinese labs. How the second squares with export rules isn't something either short summary answers, and it'll be interesting to see who asks.

53. Lenar Kess 00:16:49

In Washington, Time reports that a Senate investigation led by Senators Warren, Van Hollen, and Blumenthal says some hyperscalers misled the public about the costs and benefits of AI data centers. I only have the headline, so I'll stop there. And Wired reports that dozens of staff at Harper Collins, Simon and Schuster, and Hachette say their publishers are using AI for back-cover copy and cover art, without author consent.

54. Damra Vol 00:17:14

The Wired story is about consent. Writers sign contracts expecting a person to write the flap copy and pick the cover artist. When it turns out a model did part of that job, nobody lied exactly, but nobody asked either. I'd expect agents and authors' groups to start asking for clauses.

55. Lenar Kess 00:17:31

Last in this group, a reverse-engineering tool called REA hit 459 points on Hacker News. And a Reddit post shares a video of GLM 5.3 reverse-engineering a Dell monitor through Ghidra, connected using the Model Context Protocol. The post presents it as a vulnerability researcher with ten years in the field watching the model do the job.

56. Damra Vol 00:17:53

I like the setup. The capability comes from a general model plus a real disassembler it can drive, instead of a special-purpose security model. The model's reach into the tool is what makes the difference. And the same setup works for defenders reading firmware and for people hunting for holes to exploit.

57. Lenar Kess 00:18:10

Three talks from the AI Engineer channel this week. First, Dom, a developer experience engineer at OpenAI, presented the Codex App Server. It's an open-source protocol under the Apache 2 license, and it wraps the Codex harness. Under the hood it's a JSON-style remote-call protocol with over 120 client messages, and they cover initialization and thread management, goal configuration and file operations, and plugin discovery.

58. Damra Vol 00:18:36

The contrast he draws is with the Agent Client Protocol, which is generic. The app server gives you finer control: per-thread model selection, sandbox network permissions, and file access restrictions. That network setting is the same knob Anthropic just turned off for all of its evals. Here it's a per-thread option a client can set.

59. Lenar Kess 00:18:56

It also works with any model provider that has a compatible responses API, including LM Studio and Ollama, and a client can sign in with a ChatGPT account to use someone's existing subscription. It already powers the Codex app and the IDE extensions, plus third-party plugins in tools like JetBrains and Xcode. You connect over standard input and output or WebSockets, and there's a recently released Python SDK.

60. Damra Vol 00:19:21

A hundred and twenty messages is a big surface. If I built on it, I'd pin the harness version and treat upgrades the way I treat a database migration. With a protocol like this, stability decides whether the third-party plugins survive the next release.

61. Lenar Kess 00:19:36

Second talk, from Charles, a research scientist at Snorkel AI. In a simulated financial environment, a 4 billion parameter Qwen model reached nearly 60 percent pass-at-one on the FinQA benchmark, ahead of its 235 billion parameter sibling at 51 percent. The team built the questions from SEC filings, generated about 6,900 SQL tables, and checked question-answer pairs with programmatic tests, agent reviews, and expert verification.

62. Damra Vol 00:20:06

I trust the diagnosis. The failures in the big models were schema hallucination, context flooding from sloppy queries like select-star, and poor error recovery, where a model repeated a failed strategy instead of changing it. All three are tool-use discipline, not reasoning depth, and the ablations in the talk say the same thing.

63. Lenar Kess 00:20:25

The training recipe is modest. They used a reinforcement learning algorithm called GRPO with a binary correctness reward, ran 1,000 concurrent environments on eight H100 GPUs, and spent under 500 dollars. More complicated reward structures gave diminishing returns, and training only on simple single-table data gave the biggest lift and carried over to multi-table questions. The scripts and synthetic data are released.

64. Damra Vol 00:20:50

My caution is the scale of the test. The held-out set is 290 questions, built by the same pipeline as the training data, and graded by a small model comparing against reference answers. That's a decent result for a narrow task, and it's not evidence that 4 billion parameters beats 235 billion in

general. The speaker's own claim is enterprise reliability and domain specialization, and that's what the numbers support.

65. Lenar Kess 00:21:17

Third, two memory talks that line up. Gil Feyg is the co-founder and CTO of Merge, and he describes a context graph for agents. It mixes skills and memory with live connections, synced local copies of third-party data, and derived summaries. He argues every retrieved item should carry its provenance: the source system and fetch timestamp, the access scope and permissions, and the transformation history and invalidation rules.

66. Damra Vol 00:21:42

An Oracle developer advocate named Anders gives a four-part version of agent memory: formation, recall, evolution, and evaluation, with redaction of personal data and secrets before anything gets stored. Put the two together and what you want is a record that says who wrote this, when, under what permissions, and for which task. That sort of label on an invented tip would have made it trivial to filter.

67. Lenar Kess 00:22:08

To close, a 33-minute Latent Space conversation on protein structure prediction and data scaling. The speakers say AlphaFold 2 was a conceptual milestone because it reproduced structures deposited in the Protein Data Bank, but that protein dynamics, ground states, and structural distributions are still unresolved. Proteins are disordered and depend on context, so a static block isn't a good representation.

68. Damra Vol 00:22:32

What I took from it is that they reject the choice between craft and scale. Inductive bias from biophysics let AlphaFold 2 get a lot out of existing data instead of requiring a massive new collection. But scaling takes real engineering, like changing transformer architectures and building infrastructure for internet-scale data. Finding the right data statistics, they say, matters as much as the architecture.

69. Lenar Kess 00:22:55

They also describe a problem-first method: define the scientific goal, then decide whether the bottleneck is modeling, compute, or data. For cell genomics they say current data is nowhere near enough for something like a virtual cell, so the work is targeted data collection. Five years after AlphaFold 2, they say function, dynamics, and de novo design are still wide open.

70. Damra Vol 00:23:17

That lines up with the Snorkel talk more than I expected. In both cases somebody picks the objective first, then builds the data to match it, and the model size comes after. Snorkel built a synthetic financial world to fit the task. The protein people are asking what measurements are missing.

71. Lenar Kess 00:23:34

And it's the reverse of what happened in Philadelphia. There, the environment was wider than the task: an agent told to write example content was sitting in front of a real police form. The Snorkel team and the protein researchers spend most of their effort making the environment exactly as wide as the question. Anthropic hasn't said whether its replacement for live internet access is that narrow, and that's what its next write-up should answer. I'm Lenar.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic