One Office, Thousands of Investigations

2026-09-11 / 00:20:58

> *“Banning the account ended Anthropic's involvement. It didn't end the surveillance program.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Anthropic's September threat report describes a single Chinese intelligence office producing thousands of investigations a month, and a lone consultant in Mali building dossiers from mobile-operator data. OpenAI told employees it may slow down, then asked Congress whether an industry-wide slowdown is legal. Plus the Agents API, four agent-security preprints, and a Pentagon loan to a cloud provider.

- Axios on the surveillance cases in Anthropic's threat report
- Wired: OpenAI asked members of Congress about antitrust and a coordinated slowdown
- Axios on Congress's four remaining session days
- WSJ: Anthropic says Chinese firms routed queries through transfer stations
- Armin Ronacher on rebuilding the same abstraction again
- MMPIBench: multimodal prompt injection across six frameworks
- AgentHijack: 20.3% end-to-end success across five backends
- The Pentagon's talks to lend roughly $5B to Fluidstack

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Anthropic put out a threat intelligence report yesterday covering December through August, and the case I'd start with is a single consultant in Mali. He was pulling data out of a mobile operator and using Claude to turn it into dossiers on individual people. Not a state agency with its own cyber directorate. One contractor with an account. That's the bottom of the range in this report, and the top of it is national programs.

   - axios.com
   - axios.com
   - aljazeera.com
   - techmeme.com
   - techmeme.com
   - cnbc.com
   - cnbc.com
   - axios.com
   - techmeme.com
   - theguardian.com
   - techmeme.com
   - cnbc.com
   - techmeme.com
   - youtube.com

- lucumr.pocoo.org
- techmeme.com
- arxiv.org
- arxiv.org
- arxiv.org
- arxiv.org
- cnbc.com
- techmeme.com
- techmeme.com
- techmeme.com
- techmeme.com
- techmeme.com
- arxiv.org

2. Damra Vol 00:00:29

   The Chinese religious-affairs case is the one that changes the register for me. Anthropic describes an intelligence office running an operation that would normally need many teams of analysts, and says that single office was producing thousands of investigations a month. Thousands, every month. That isn't an analyst going faster with a better tool. That's one office with the throughput of a department, and the report puts the difference down to the model.

3. Lenar Kess 00:00:55

   Sam Sabin at Axios went through the surveillance cases, and the targeting detail is what sticks. The systems were rating how politically sensitive a given post was, and then selecting people out of that pool for what the report calls control.

4. Damra Vol 00:01:10

   And Anthropic defines control in the document. It means coercive questioning, or closer monitoring of someone's movements and communications. So the pipeline runs all the way from a social media post to a person being pulled in for a conversation they didn't want to have. The model sits in the middle, and the ranking is its job.

5. Lenar Kess 00:01:30

   Let me lay out where we're going today. We start here with the threat report. Then OpenAI, which told employees it's considering slowing down cutting-edge development and then asked members of Congress whether an industry-wide slowdown would even be legal. Then Washington's reaction, which has four session days to happen in. Then Anthropic's distillation allegations against Chinese

labs. Then OpenAI's new Agents API. Then a stack of security preprints on agent attacks. The last stretch is capital, where the strangest item is the Pentagon considering a loan to a cloud provider.

6. Damra Vol 00:02:05

   Back in the report for a second, because the non-surveillance cases matter too. There's an Iranian operation running a malicious Firefox extension that harvested identity information. There are five instances Anthropic disrupted involving biological work, two of which it classifies as gain-of-function. And Al Jazeera picked up a claim about missile guidance software in Yemen.

7. Lenar Kess 00:02:28

   The biological section has the narrowest sentence in the whole document. Anthropic writes, quote: We take these cases as evidence not of the imminence of biological threats currently uplifted by Claude, but rather as evidence that significant dual-use research efforts are associated with state actors of concern who routinely evade our access controls. End quote.

8. Damra Vol 00:02:49

   That sentence claims something much smaller than the headlines it produced. They aren't saying Claude uplifted anyone's weapons program. They're saying that state actors interested in dual-use research keep getting through their access controls. Those are different claims, and the second one is about account security rather than model capability.

9. Lenar Kess 00:03:09

   One more detail I haven't seen picked up much. Every surveillance case they describe used Haiku, Sonnet, or Opus. None of them used the newer Fable or Mythos class models.

10. Damra Vol 00:03:20

    Which tracks with something Jacob Klein said. When enforcement creates friction, operators move to open-source models. The Mali platform ended up deployed locally against other models. So banning the account ended Anthropic's involvement. It didn't end the surveillance program.

11. Lenar Kess 00:03:38

    Klein's line on the overall picture was blunt. He said, quote: Authoritarian states are using AI for surveillance, repression and influence operations today. It's no longer hypothetical. End quote.

12. Damra Vol 00:03:51

    And the caveat has to ride along with all of it. This is Anthropic's account of Anthropic's own enforcement. Every case in here came through their API and got caught. The operations that never

touched Claude, or touched it and went undetected, aren't in the document. It's a substantial report and it's also a partial view by construction.

13. Lenar Kess 00:04:12

    Bloomberg reported that Sam Altman told OpenAI employees the company is considering slowing down cutting-edge development. And then Maxwell Zeff at Wired reported the follow-on, which is that OpenAI has been asking members of Congress whether orchestrating an industry-wide slowdown would be legal under antitrust law.

14. Damra Vol 00:04:31

    That second half is the more interesting document. A company doesn't ask Congress whether coordination is legal unless it has thought about coordinating. And the answer under the Sherman Act isn't encouraging for them, because competitors agreeing to restrict output is close to the textbook description of what the statute prohibits. It doesn't matter much that the output in question is model capability rather than barrels of oil.

15. Lenar Kess 00:04:57

    And there's no safety exemption written into it. The usual route for something like this is a statutory carve-out, which Congress would have to pass, or a regulator setting the limit so that it isn't the companies agreeing among themselves.

16. Damra Vol 00:05:10

    Then there's who isn't in the room. Even if OpenAI, Anthropic, and Google all agreed to hold back, that agreement doesn't bind open-weight releases and it doesn't bind Chinese labs. So the coordination that's hardest to arrange legally is also the coordination that covers the smallest fraction of the frontier.

17. Lenar Kess 00:05:29

    CNBC's reporting put more of the motive on the record. Altman talked about self-improving systems, and described concerns that read as existential rather than commercial.

18. Damra Vol 00:05:39

    The reception at Y Combinator's Demo Day wasn't sympathetic. Garry Tan's advice on what to do about all of this was, in short, do nothing. Coming from the head of the largest startup accelerator, that's a position about who bears the cost of a slowdown. OpenAI isn't the company that gets squeezed first if the frontier freezes. The companies building on top of the next model are.

19. Lenar Kess 00:06:03

The tension there is plain. The lab with the most to lose from competition is the one proposing that everyone compete less. That doesn't make the safety concern insincere, and Altman may believe every word of it. But the structure of the proposal happens to favor the proposer, and Congress will notice that before it notices the argument.

20. Damra Vol 00:06:22

    There's also a timing question. Bloomberg says considering. No slowdown has been announced, no capability ceiling has been named, and no date has been given. Right now this is a conversation with employees plus a legal question put to legislators. Until one of those becomes a policy with a number attached, it tells you what OpenAI is worried about rather than what OpenAI will ship.

21. Lenar Kess 00:06:46

    That's about the right size for it today. The next development that would change my read is a specific commitment with a date on it, or a member of Congress saying publicly what answer they gave.

22. Lenar Kess 00:06:57

    Washington's reaction is the other half of this, and it has a clock attached. The House is in session four more days before Election Day. That's the window for anything legislative.

23. Damra Vol 00:07:07

    And it got onto the agenda through Josh Coxon's post, which Axios says has now passed a hundred and fifty million views. A resignation post from one person outran every white paper ever written on the subject.

24. Lenar Kess 00:07:20

    Ted Lieu has a bipartisan bill for a human-activated kill switch on advanced systems. Bernie Sanders and Greg Casar have a different one that would pause development and ban work toward superintelligence. Ruben Gallego wants a Select Committee on AI. And Speaker Johnson's own bill is narrow, because it's about data centers.

25. Damra Vol 00:07:40

    That spread tells you there's no consensus on what the problem even is. A kill switch is a technical control. A development pause is industrial policy. A select committee is a jurisdiction fight. Those three bills don't disagree about the answer so much as about which question Congress is answering.

26. Lenar Kess 00:07:58

Then there's the anonymous ranking member Axios quoted. On Coxon's warning: I think he's wrong. On the idea of having staff learn about AI: it would be a waste.

27. Damra Vol 00:08:09

That's a ranking member. Someone senior enough to shape what their committee does with the issue, saying staff education on it isn't worth the hours. Whatever you make of the substance of Coxon's claims, institutional capacity is a separate question and that quote answers it.

28. Lenar Kess 00:08:25

The President's position is on the record now too. Bloomberg and CNBC both have him brushing aside the extinction warnings, saying the United States leads China by about a year, and saying that without that lead the country would be in a very bad position.

29. Damra Vol 00:08:40

Which turns the whole thing into a race, and a race doesn't have a brake pedal in it. If the operating assumption is a one-year lead over China, then every argument for slowing down arrives sounding like an argument for giving up the lead.

30. Lenar Kess 00:08:54

There's a concrete gap in the rules right now. The European Union's AI Act already requires serious-incident reporting. The Trump administration's framework has no equivalent requirement.

31. Damra Vol 00:09:05

So if an American lab has an incident of the kind Coxon was describing, no law obliges them to tell anybody. That's a missing reporting line, and it's small enough to pass in four session days if anyone wanted it to.

32. Lenar Kess 00:09:18

Gaby Hinsliff wrote in the Guardian about, quote, an industry now practically begging to be saved from itself. Which is roughly where we are. OpenAI asking Congress whether coordination is legal, and Congress short on staff time to learn the subject.

33. Lenar Kess 00:09:34

Anthropic also told the Wall Street Journal that Chinese companies have been distilling its models. They describe thousands of fake accounts and millions of queries from real users, all routed through what Anthropic calls transfer stations outside China.

34. Damra Vol 00:09:48

    Transfer stations is a tidy piece of vocabulary for a dull mechanism. It's intermediary infrastructure outside Chinese jurisdiction that makes the traffic look like it originates somewhere else. Anthropic can act on the fake accounts. The millions of real user queries are much harder, because those are real customers doing real work, and some fraction of the output ends up as training data somewhere else.

35. Lenar Kess 00:10:12

    CNBC's version names Alibaba alongside DeepSeek. And the timing is notable, because this comes a day after the National Security Agency, CISA, and the FBI jointly named six Chinese AI firms in an advisory.

36. Damra Vol 00:10:27

    We went through that advisory earlier this week, so the new increment is a lab saying publicly what the agencies said. The revenue number sitting next to it is what separates this from the version we discussed back in June.

37. Lenar Kess 00:10:40

    Moonshot. Their annualized recurring revenue was three hundred million dollars in June. By August it had reached one billion. They're targeting two billion by the end of 2026, all of it driven by Kimi K3.

38. Damra Vol 00:10:55

    So the distillation complaint isn't a research curiosity anymore. It's about a competitor tripling revenue in two months on a model that the American labs say was partly built out of their own outputs. That's a commercial injury claim wearing a national security coat.

39. Lenar Kess 00:11:10

    Though I'd hold the causal link loosely. Nobody has shown that Kimi K3's revenue depends on distilled data. Anthropic alleges distillation, Moonshot has revenue, and connecting the two is an inference. It's Anthropic's inference.

40. Damra Vol 00:11:25

    Fair. And it puts Garry Tan's do-nothing advice back in view from a different angle. If the Chinese labs are compounding at that rate, a voluntary American slowdown has a very specific price, and the people arguing for it have to say what they think that price is.

41. Lenar Kess 00:11:42

    OpenAI shipped an Agents API, and the shortest description is that they've taken the Codex harness and hosted it. The orchestration loop, session management, and context compaction all move server-side.

42. Damra Vol 00:11:54

    The division of labor is the interesting bit. OpenAI keeps the orchestration and hands you the execution sandbox, which can be theirs, a third party's, or your own. So the product they're claiming is the agent loop, which is exactly what everyone has been rebuilding in-house for two years, and they're leaving alone the place where code actually runs.

43. Lenar Kess 00:12:15

    Tools come in over Model Context Protocol. Skills are runbooks. And there's programmatic tool calling, which is the feature I'd point at. Instead of dumping a log file into the context window, the agent writes code to filter it, and only the result comes back.

44. Damra Vol 00:12:31

    That's a real cost line. Filtering logs inside the context window is where token budgets go to die. And the demo emits a structured incident findings file at the end, which tells you the customer they have in mind is an on-call engineer rather than a chatbot user.

45. Lenar Kess 00:12:47

    The counterpoint got more traction on Hacker News than the launch did. Armin Ronacher's post, titled Astra for Coding: Why Are We Doing This Again?, was sitting at two hundred and ninety-five points.

46. Damra Vol 00:12:59

    His objection is roughly that we keep rebuilding the same abstraction with a new vendor name on it. That's not a small point when the vendor in question owns the loop. If your orchestration lives inside OpenAI's API, switching models later stops being a configuration change.

47. Lenar Kess 00:13:16

    There's one caveat I'd put on the whole segment. The only source for what this thing does is OpenAI's own launch video, which runs two minutes and eighteen seconds. No independent evaluation, no pricing analysis, and nobody outside the company has run it under load.

48. Damra Vol 00:13:32

Related, and a little sideways. The Y Combinator spring 2026 batch is four percent native-mobile-first. Back in 2013 that number was fifteen percent. Whatever founders believe the next platform is, they aren't betting on phones, and a hosted agent loop is a reasonable guess at what they're betting on instead.

49. Lenar Kess 00:13:53

    Four security preprints arrived at once, all version one and none peer reviewed, and together they make agent attacks look more specific than the headline numbers suggest.

50. Damra Vol 00:14:03

    Start with MMPIBench, because it has the most structure behind it. Seven hundred and twenty runs in total. They span six agent frameworks and five models. The injected instruction arrives through six different visual carriers. Attacks were attempted in twelve point eight percent of cells, and completed in about one percent.

51. Lenar Kess 00:14:24

    And the distance between attempted and completed closes almost entirely at the planning step. The agent reads the injection, starts to act on it, and then something in planning drops it.

52. Damra Vol 00:14:34

    The per-model spread is much wider than the average. One model never attempts the injected instruction at all, and recognizes it as an injection fifty-nine point seven percent of the time. Two others attempt it around twenty-three point six percent of the time. So multimodal prompt injection risk isn't one number. It's a property of the specific model you deployed.

53. Lenar Kess 00:14:57

    The audio result is what changed my priors. Just two of the five models will ingest audio at all. Of the six frameworks, only three will deliver it. But where audio does arrive, attacks complete in forty-nine percent of cells. For one model it's seventy-five.

54. Damra Vol 00:15:14

    So audio isn't safe. Audio is mostly unplumbed. The low aggregate number is an artifact of the pipeline, and the moment more frameworks wire up audio input, that number moves a long way. What's holding it down is missing plumbing, and plumbing gets built.

55. Lenar Kess 00:15:31

Second paper is AgentHijack. Six hundred instance-level online cases across five backends. The number that got quoted around is eighty-four point five percent target attack success. The number that belongs next to it is twenty point three percent end to end.

56. Damra Vol 00:15:47

    And the middle figure explains the gap, which is forty-seven percent action parse. So the attack gets the model to name the malicious target most of the time, but converting that into a parsed action and then a completed chain drops it by a factor of four. Quoting the eighty-four on its own would be misleading.

57. Lenar Kess 00:16:06

    There's a detail in the trajectories I hadn't seen described before. In some runs the agent executes the malicious terminal command, and then goes back and finishes the benign task it was originally given.

58. Damra Vol 00:16:17

    Which is why detection is so hard here. The user-visible outcome is success. The task completes, the output looks right, and the compromise is a side effect nobody reviews. Any monitoring that keys on task failure sees nothing at all.

59. Lenar Kess 00:16:32

    Third one is the cipher attack paper. No fine-tuning involved, because the model learns the cipher through prompting and in-context learning, and then the harmful payload arrives looking like gibberish, so the harmfulness classifiers never fire on it. They demonstrate it against Anthropic, Google, and OpenAI models.

60. Damra Vol 00:16:51

    That's a structural problem for input filtering as a category. If the model can be taught an encoding at inference time, then any filter inspecting plaintext is inspecting the wrong layer. And the fourth paper, AgentAudit, goes the other direction, scoring ten dimensions over the full execution trace rather than the final output. Claude Sonnet 5 comes in at ninety-five point one composite trust. GPT-5 sits at eighty point six.

61. Lenar Kess 00:17:19

    With a limitation the authors flag themselves, which I appreciate. The judge model was one of the models being evaluated.

62. Damra Vol 00:17:26

You can't prompt your way out of that one. It's the same circularity everybody hits when the evaluator and the evaluated come from the same family. Still, scoring the trace instead of the answer is the right direction, and it's the direction the attack papers point at, because the compromise in AgentHijack happens in the trace and never shows up in the answer.

63. Lenar Kess 00:17:47

    Capital roundup to close. First, Oracle reported thirty percent revenue growth, with cloud infrastructure up a hundred and twenty-one percent year over year. The stock rose six percent on it.

64. Damra Vol 00:17:59

    That's the demand side confirmed on an income statement, which is more than most of this week's AI infrastructure news offers.

65. Lenar Kess 00:18:06

    Then the odd one. The Pentagon is in talks to lend roughly five billion dollars to a neocloud called Fluidstack, with Palmer Luckey's Erebor Bank advising.

66. Damra Vol 00:18:16

    Government financing compute rather than buying it. That's a different instrument with different consequences, because a purchase gets you capacity while a loan gets you a creditor's position and some leverage over who else the borrower serves. And Erebor advising on a Defense Department facility to an AI cloud is a small circle of people.

67. Lenar Kess 00:18:36

    In Shanghai, Enflame closed up a hundred and eighty-eight percent on its debut. It raised about nine hundred and ten million dollars and finished at a market capitalization of twenty-six point three billion. Two new billionaires came out of it.

68. Damra Vol 00:18:51

    Domestic Chinese accelerator demand, priced. That's the other half of the export-control conversation and it doesn't get said much. Restricting Nvidia creates a market, and that market now has a public comparable attached to it.

69. Lenar Kess 00:19:05

    Ayar Labs raised a hundred and fifty million dollar extension to its March Series E, for optical chip-to-chip interconnect. And Fidji Simo joined Nscale's board, recruited by Sheryl Sandberg, while staying on as a part-time OpenAI advisor.

70. Damra Vol 00:19:20

    Two items about the same bottleneck from opposite ends. Ayar is solving it in silicon, and Simo's board seat puts an OpenAI-adjacent person inside a compute provider. Both are bets that the constraint stays physical for a while yet.

71. Lenar Kess 00:19:35

    Two more papers. There's a cyber-financial contagion model built on sixty vendors and two hundred and twenty banks. The graph carries about twenty-five hundred vendor-to-bank edges and fourteen hundred interbank exposures. Patch latency dominates the outcome.

72. Damra Vol 00:19:51

    And the data is entirely synthetic, which the authors say outright. So it's a model of a contagion mechanism rather than a measurement of one. The patch-latency finding is a property of their graph. It would be a result if somebody ran it on actual vendor relationships, and nobody has.

73. Lenar Kess 00:20:08

    And OpenAI pulled its sponsorship of the Caltech math hackathon after criticism about slop mathematics. We went through the authorship dispute earlier this week. The new piece is that the money moved.

74. Damra Vol 00:20:21

    Withdrawing a sponsorship is the cheapest available response, and it's also the clearest signal of whose approval OpenAI thinks it needs. Mathematicians are the audience for the proof claims. If that community decides the output is noise, the claims stop working as evidence.

75. Lenar Kess 00:20:37

    Anthropic's threat report is public, all four preprints are on arXiv, and the House has four session days left. If a serious-incident reporting requirement shows up inside one of those bills, that's the piece of this week that would change what labs are obliged to tell anyone. Thanks, Damra. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic