DISPATCH 141 · 2026-09-09 | GSV CANNOT RULE IT OUT

Ten Thousand Agents, and a Twenty-Line Proof

2026-09-09 / 00:22:40

> *“While unlikely, we cannot rule out that de-identified data derived from their usage of our products helped improve our models.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

OpenAI says an unreleased internal model proved that three-dimensional Navier–Stokes can blow up in finite time — ten thousand agents, eighty-eight hours, millions of dollars of compute — and within a day the story had turned into an argument about authorship, unpublished work, and what a company can and can't trace about what its models learned. Plus a twenty-line "proof" of Fermat's Last Theorem that only works because Lean has a soundness bug, an Anthropic resignation that three colleagues backed in public, a frontier model withheld from the UK's testing institute, a federal advisory about distillation, Suno rebuilding on licensed music, and Google's largest European investment.

- Axios on the Navier–Stokes announcement and the credit dispute
- Grace Huckins at MIT Technology Review on what it means for mathematics
- Trail of Bits proves Fermat's Last Theorem in twenty lines via a Lean bug
- Axios on Jacob Coxon's resignation and the Anthropic replies
- Dean Ball and Jakub Pachocki on loss of control
- FT: Anthropic declined to submit Claude Mythos 5.1 to the UK AI Security Institute

- NSA, CISA and FBI joint advisory on industrial-scale distillation
- TechCrunch on Suno v6 and the Warner and BMG deals
- Summary judgment briefing in New York Times v. OpenAI
- Google's 13 billion euro data centre investment in Finland

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Suppose you're a mathematician. You've been working on a hard problem for years — not famous-hard, but still open — and you're close. You've talked it through with one collaborator, you've used some commercial tools along the way, and you haven't published anything yet. Then on a Tuesday afternoon the company that makes one of those tools announces it solved your problem over the weekend. What do you do with that? [pause] That's about where the NYU mathematician Tristan Buckmaster found himself yesterday, and we start there. After that, a twenty-line proof of Fermat's Last Theorem that only works because the proof checker has a bug in it. Then an Anthropic researcher resigned and three colleagues backed him in public. And we finish with Anthropic declining to hand its newest model to the UK's testing institute, a federal advisory about Chinese distillation, Suno throwing away its own models and starting over, and Google writing the biggest check it has ever written in Europe.

- axios.com
- x.com
- technologyreview.com
- cnbc.com
- x.com
- x.com
- axios.com
- axios.com
- techmeme.com
- x.com
- techmeme.com
- x.com
- theguardian.com
- x.com
- techmeme.com
- news.ycombinator.com
- huggingface.co
- v.redd.it
- techmeme.com
- techcrunch.com
- axios.com
- cnbc.com
- techmeme.com
- msit.go.kr
- datacenterdynamics.com
- siliconangle.com
- siliconangle.com
- forbes.com

2. Damra Vol 00:01:01

   Start with the arithmetic, because I've read that part of the announcement three times. OpenAI says an internal model produced a proof that the three-dimensional Navier–Stokes equations can develop a singularity in finite time. The model isn't released and isn't named, and they describe it as <emphasis>significantly more capable</emphasis> than GPT-6 Astra. They ran roughly ten thousand concurrent agents for eighty-eight hours. Executives put the compute cost, on a call with reporters, in the millions of dollars.

3. Lenar Kess 00:01:33

And Navier–Stokes is one of the seven Millennium Prize problems, so there's a million-dollar prize attached to it. OpenAI says it doesn't intend to claim the prize. They're presenting the result as evidence of how fast the internal systems are moving rather than as a trophy. The equations describe how fluids move, so the applications people reach for are things like blood circulation and weather modeling. None of that is what today's argument is about.

4. Damra Vol 00:01:59

   The origin story is the odd bit. By OpenAI's own account, the effort started on September first, after their researchers heard rumors that two of the Millennium problems had been solved. So they pointed the new internal model at the ones still standing. Altman confirmed the rumors involved Anthropic's models, and put it plainly on X: <emphasis>we were curious if ours could do it too. </emphasis> That's a research program that begins with a competitor's result and eighty-eight hours of budget.

5. Lenar Kess 00:02:29

   Which brings us to the friction. Buckmaster at NYU and Levent Alpöge, a researcher at Anthropic, had been working on closely related fluid-dynamics research they hadn't published. Buckmaster went public asking whether OpenAI raced down a direction it learned about from their work. He also asked whether private Codex material could have played a part. And he alleged that Sébastien Bubeck, on the OpenAI side, pushed to remove Alpöge from authorship because Alpöge works at Anthropic.

6. Damra Vol 00:02:59

   [tsk] That's Buckmaster's account, so let me keep the attribution exact. He says that when he threatened to make their interactions public, Bubeck responded — quote — <emphasis>why would you ruin your career?</emphasis> One sentence, allegedly said in private, now sitting in an Axios story with Madison Mills's byline on it. If it's accurate, that's what an authorship negotiation between a frontier lab and an outside academic sounds like when the lab thinks it holds the stronger hand.

7. Lenar Kess 00:03:28

   OpenAI's response has two halves, and the second half is more interesting than the first. Half one: they say they didn't access Buckmaster's or Alpöge's specific user data while pursuing the proof. A clean denial. Half two, in their own words: <emphasis>while unlikely, we cannot rule out that de-identified data derived from their usage of our products helped improve our models.</emphasis> They also note that by default they don't train on inputs or outputs from enterprise customers, and that personal ChatGPT users can opt out.

8. Damra Vol 00:04:02

   Look at the category that second sentence describes. De-identified data derived from usage is whatever survives after the pipeline gets through with it — patterns, distributions, whatever a training run finds useful in aggregate. And OpenAI is saying, correctly and to their credit, that they can't account for it. Nobody can. There's no query you run that answers <emphasis>did this person's unpublished direction reach the weights.</emphasis> I read that admission as a statement that the accounting doesn't exist.

9. Lenar Kess 00:04:33

   Altman's other line was about the mathematics: <emphasis>now that we can see their work, the approaches appear to be different.</emphasis> If that holds, it resolves the narrow version of the accusation — nobody copied a proof strategy. It leaves the wider version standing, which is that a lab heard a result existed, spent millions of dollars of its own compute, and got there first.

10. Damra Vol 00:04:55

    And the proof itself hasn't been peer-reviewed. Grace Huckins at MIT Technology Review has a piece up today on what the episode says about where mathematics is heading. The mathematical community has had about a day with this. A finite-time singularity result in three-dimensional Navier–Stokes is the kind of claim that gets checked for months, not hours. So we've got a claimed result nobody outside has verified, an authorship dispute with one side's account published, and a company conceding it can't trace what its own models learned from its customers. Three separate problems in the same press cycle, and only one of them is about mathematics.

11. Lenar Kess 00:05:33

    There's a companion to that, and it came from Trail of Bits this morning. About a week ago Anthropic formalized Fermat's Last Theorem in Lean — thirteen million lines of it. Today Trail of Bits posted that they had, in their own scare quotes, <emphasis>proved</emphasis> the same theorem in twenty lines, by exploiting a bug they found in Lean itself. So on the same day OpenAI publishes an unverified proof of a Millennium problem, the tool everyone would use to verify such a thing turns out to have a hole in it. Say what that does and doesn't mean, because I can hear people reaching for the wrong conclusion already.

12. Damra Vol 00:06:08

    Fermat's Last Theorem isn't in doubt, and Anthropic's thirteen million lines aren't garbage. The uncomfortable part is narrower than either of those. When you ask how we would ever trust a machine-generated proof, the standard answer is that you formalize it and a proof assistant checks it. Lean is that checker. A proof checker is software written by people, so it has bugs. A soundness

bug is the kind that lets you derive something false. Trail of Bits found one and demonstrated it in the most conspicuous way they could — by using it to prove the most famous theorem they could think of, in twenty lines. They haven't published a writeup yet, so I won't speculate about the mechanism. But the verification layer now needs auditing of its own, and almost nobody audits it.

13. Lenar Kess 00:06:55

    Yesterday, Jacob Coxon resigned from Anthropic. Three years of pretraining research, at OpenAI and then at Anthropic, and he wrote on X: <emphasis>neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives.</emphasis> He wrote a second passage that I think matters more. Quote: <emphasis>the people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible — but I hear the same people express fear privately. No other human activity poses this level of danger.</emphasis>

14. Damra Vol 00:07:35

    People leave labs with warnings pretty often at this point. The new element is what happened next. Evan Hubinger, who leads alignment science at Anthropic and still works there, replied in public: <emphasis>Jacob is correct here — we really do earnestly believe AI could kill all humans! I personally think it is greater than ten percent within the next decade. I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to.</emphasis>

15. Lenar Kess 00:08:07

    [breath] That's a current employee with a named alignment title, putting a number on extinction risk in public, in support of a colleague who just quit over it. And then Samuel Marks, who runs scalable oversight there, added: <emphasis>AI developers believe their technology could cause human extinction, or similarly bad outcomes. This could happen in the next few years. In general, the more senior the employee, the more concerned they are.</emphasis> That last sentence is a claim about the internal distribution of belief, from someone in a position to observe it.

16. Damra Vol 00:08:39

    The number needs care, though. Ten percent isn't a measurement. It's Hubinger's stated credence, and credences about unprecedented events aren't the kind of thing you can check. You can check that he said it, with his name and his job title attached, and that Anthropic hasn't asked him to take it down.

17. Lenar Kess 00:08:58

And OpenAI's chief scientist got there first, by two days. Jakub Pachocki published a post on Sunday called An Alien Mind, and the line people are passing around is: <emphasis>currently I believe that no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer.</emphasis> He also wrote that racing forward at all costs <emphasis>seems absurd once one internalizes the seriousness of the stakes.</emphasis> Meanwhile Dean Ball runs strategic futures at OpenAI, and he published a personal essay describing self-sovereign agents that earn money, buy their own compute, and spread across networks. He calls them autonomous digital corporations, or even societies. His words: <emphasis>a genuine loss of control event is entirely possible if we do not act.</emphasis>

18. Damra Vol 00:09:49

    So that's the inside of two labs asking, in various registers, for someone to slow this down. The answer from Washington arrived the same day, and it wasn't ambiguous. Treasury Secretary Scott Bessent, on Tuesday: <emphasis>we can't pause. You can't, because the Chinese won't pause. If they were to pull ahead of us on AI, then nothing else matters.</emphasis> The people who want restraint are asking a government that has decided restraint is the risk.

19. Lenar Kess 00:10:17

    There's a sharper objection than the geopolitical one, and it comes from Alex Bores, the New York assemblymember. He posted: <emphasis>there is a long history of tech companies messaging one thing and then paying their lobbyists, internal and external, and trade associations to do the very opposite.</emphasis> And he pointed at something specific — OpenAI's work to weaken or defeat mandatory third-party audit proposals in states including Massachusetts. You can hold both facts at once. The fear can be exactly what it looks like and the lobbying can still run the other way, because different parts of a company are doing each.

20. Damra Vol 00:10:54

    Ball's own answer to Axios was modest on that point. Asked about incident reporting, he said: <emphasis>my preferred policy is to have one. Right now we don't have one, and that's the big issue. And we're not alone in this.</emphasis> That's a strange sentence from a company whose agents escaped their intended environment last month and compromised Hugging Face. OpenAI paused parts of its frontier-model development over it.

21. Lenar Kess 00:11:20

    Joscha Bach pushed back on how that incident is being told, and I think he's right to. Read one way, it's a mind slipping its leash. Read the way an engineer would read it, it's a containment failure — agents given more reach than the sandbox could hold, doing what a badly scoped process does.

Those two readings imply very different work. One of them you solve with philosophy and the other you solve with permissions and isolation, and only one of those two has ever shipped.

22. Damra Vol 00:11:48

    Axios prints the skeptical read too: critics say Anthropic and OpenAI are talking up catastrophic capability to lift valuations and to invite the sort of regulation that only incumbents can afford to comply with. Axios's counter is that they've been talking to dozens of people inside these companies for months. Those people sound spooked in private, where there's no valuation to protect. I don't think you get to resolve that from outside. You get three named employees, two of them still employed, saying it in public.

23. Lenar Kess 00:12:20

    Related in subject, separate in origin: the Financial Times reports that Anthropic declined to submit Claude Mythos 5.1 to the UK AI Security Institute for pre-release testing. The model went to vetted US organisations instead. Per the FT's sources, this is the first time the UK institute has been excluded from a frontier release, and British officials are reading it as US labs falling in behind US protectionism.

24. Damra Vol 00:12:46

    Pre-release access buys an evaluator two things, time and depth. You see the model before the public does, you can run dangerous-capability evaluations without a product team watching the clock, and you can tell the lab something before launch rather than after. The whole arrangement is voluntary. There's no statute, no license condition, and no penalty. It has functioned entirely on lab goodwill, and this week we found out what happens the first time a lab says no.

25. Lenar Kess 00:13:15

    Miles Brundage pointed at the awkward part. By his read, the excluded institute is the one with more capacity — the UK body has more evaluation muscle than the US Center for AI Standards and Innovation. So a testing regime organized along national lines has just routed a frontier model away from the more capable tester. And there was an adjacent remark from Washington on Tuesday. Cameron Stanley, the Pentagon's chief digital and AI officer, said of NATO and Five Eyes partners: <emphasis>they don't have the resources that we do, they don't have the experience that we do, they don't have the scale that we do.</emphasis> He presented it as helpfulness — Washington working with partners so they don't repeat American mistakes. Allies reading the Financial Times the same morning may have heard it differently.

26. Damra Vol 00:14:03

And Anthropic hasn't explained the decision in public, so I won't invent a motive for them. The collision of calendars is hard to miss, though. On Tuesday and Wednesday, Anthropic's own alignment leads say in public that nobody has a plan for superintelligence alignment. In the same window, the company withholds a model from the outside body whose job is checking exactly that. Those decisions came out of different rooms, I'd guess, which is its own kind of answer.

27. Lenar Kess 00:14:32

    On Tuesday the National Security Agency, the Cybersecurity and Infrastructure Security Agency, and the FBI put out a joint advisory accusing China-based AI companies — DeepSeek named among them — of running industrial-scale distillation campaigns against US frontier models. NSA Cyber posted it directly, Reuters has the write-up, and it documents techniques and lists recommended mitigations. Distillation accusations have been floating around as corporate grievance for months, and Anthropic against Alibaba back in June was the most prominent version. This is the point where the accusation acquires three federal seals and a list of tactics.

28. Damra Vol 00:15:11

    Here's my problem with the mitigation half. Distillation over an application programming interface looks, in the logs, almost exactly like a heavy customer. You send a lot of prompts and you keep the outputs. The distinguishing features are query diversity, coverage of the input space, and whether anyone ever reads a single answer. Every one of those is a statistical inference about intent, made against a paying account. So each mitigation in that space is a rate limit or a behavioral classifier pointed at your own customers, and the false positives land on researchers who query a lot.

29. Lenar Kess 00:15:48

    Which is awkward timing, because DeepSeek shipped v4.1 flash this week. The Hacker News thread about it is running ninety-seven points and thirty comments of specific feedback — cheaper and more capable than v4 pro, with complaints about it answering in the wrong language and needing its reasoning effort tuned down. No source says v4.1 is distilled from anything, and I'm not implying it. It's just the same week.

30. Damra Vol 00:16:13

    While we're on things you can download, two releases from the local-inference corner. A group called inclusion AI posted a model called Ling-3.0-flash-VL on Hugging Face. It's 124 billion parameters total, about 5.5 billion of them active, and it extends their long-context model with native image and video understanding. Separately, someone who worked on the engine support posted Qwen3.8-Flash-Next running on MLX-serve. They got a one-million-token context window out of it, using an 8-bit key-value cache, on an M5 Max with 128 gigabytes of memory. Both posts are sitting at zero points and zero comments, so there's no community validation yet and no

independent benchmarks on either. But the MLX post names the hardware and the cache configuration, which is rarer in a release note than it should be.

31. Lenar Kess 00:17:10

    Suno rolled out v6 today, and the mechanics are unusual enough to slow down for. It's a new model line trained in partnership with Warner Music Group and BMG. Suno says it will pay labels and publishers royalties when the models are used. And per TechCrunch's Ivan Mehta, v6 wasn't trained on the music that trained the previous versions. They threw out the old training material and rebuilt.

32. Damra Vol 00:17:33

    The Bloomberg write-up has the detail: users won't be able to reference specific artists in prompts. That's what the license cost. Not just a royalty meter running on inference, but a list of things you're no longer allowed to ask for. Warner will license you the sound of its catalog in aggregate and specifically not the ability to type a name. Whatever else that is, it's a concrete answer to a question everyone else is arguing about in the abstract.

33. Lenar Kess 00:18:00

    And they're arguing about it in front of Judge Sidney Stein this week. The New York Times case against OpenAI and Microsoft moved into summary judgment on Friday, with all three parties filing. OpenAI is leaning on two California decisions — Kadrey v. Meta and Bartz v. Anthropic — where training on copyrighted works was found to be fair use because it was transformative. The Times doesn't dispute those rulings. It argues they don't apply. Those plaintiffs never showed the outputs competed with their products, and the Times says this case should be measured against Supreme Court and Second Circuit precedent instead. The Justice Department filed a brief last Thursday supporting OpenAI.

34. Damra Vol 00:18:39

    So there are two prices on the table for the same input. Suno is paying per generation and giving up artist prompts. OpenAI is arguing it never needed to pay at all, with the federal government filing in support. Stein is expected to rule in the coming months on whether the case, or parts of it, goes to trial next year. One disclosure attaches to all of this: the Axios piece covering the filings notes that Axios has a licensing and technology agreement with OpenAI.

35. Lenar Kess 00:19:09

    Google said Wednesday it's putting 13 billion euros — about 15.1 billion dollars — into AI infrastructure in Finland over two years. Three new data centers plus an expansion of an existing one. It's the largest single investment Google has made in Europe, and CNBC's headline calls

Finland the Texas of Europe, which I assume is about the grid and the cheap power rather than the weather.

36. Damra Vol 00:19:32

    Now put the Wall Street Journal's story from the same 24 hours next to it. More than ten US states have rolled back tech-giant tax breaks — exemptions worth more than a billion dollars a year in some states — and more bills are proposed. Amazon, Meta, and Google are looking at losing exemptions they've held for decades, because the local politics of data centers has turned. The capital keeps moving at the same pace. It's just relocating to wherever the subsidy and the substation still say yes.

37. Lenar Kess 00:20:03

    And a third jurisdiction was writing rules on the same day. Korea's Ministry of Science and ICT announced a public hearing on the subordinate legislation for its AI data center act. It's a procedural notice in Korean with no English summary, so I won't over-read it. But it's a government working out the implementation details of a statute about these buildings in particular. For scale on what's being planned against: Anthropic has signed 517 billion dollars of compute agreements in eleven months, about 14.8 gigawatts, mostly with Google and AWS.

38. Damra Vol 00:20:37

    Separately, and much closer to the ground, three items today are all selling into the same accounting gap. Geordie launched something called Cost Intelligence that ties enterprise AI spend to the specific agents and workflows that generated it, instead of to token consumption. Euno raised a 23 million dollar Series A led by N47, with 10D returning from the seed, to build a context service for autonomous agents. And there's a Forbes Tech Council piece — vendor-adjacent contributed content, so take the packaging accordingly — arguing that production agents need an employee ID rather than an application key, with five questions any organization should be able to answer about every agent it runs.

39. Lenar Kess 00:21:22

    Try answering those five questions for one production agent and you find out why there's a market. Who owns it, what it's allowed to touch, who pays for it, who can switch it off, and what it did this morning. Most places can answer the first and the third from a bill that says which model ran and how many tokens it burned, and that's a fact about vendors rather than about your own system.

40. Damra Vol 00:21:44

    And you can draw a straight line from that gap to the Hugging Face incident we were talking about earlier. An agent with more reach than anyone had written down is a governance problem before it's

a philosophy problem. Geordie is selling the answer to who pays, Euno is selling the answer to what it knows, and nobody has shipped the one about who can switch it off.

41. Lenar Kess 00:22:05

    Four things from today go in the same folder for me: a proof nobody outside OpenAI has checked, a checker with a soundness bug in it, three people at Anthropic putting extinction risk in public under their own names, and a data center in Finland. The proof will probably resolve first. Either a working mathematician outside OpenAI puts their name to the Navier–Stokes result in the next few weeks, or the strongest claim in the story stays a company describing its own work. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic