The Robots Will Not Be Taking Over

2026-09-15 / 00:24:50

> *“Four labs agreed on a direction over the weekend, and by Monday afternoon nobody could agree on who gets to hold the stopwatch.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A voluntary slowdown agreed by four frontier labs over the weekend ran into two obstacles on Monday: a president who calls the premise a hoax, and an industry that can't name a single evaluator everyone would accept.

- Trump phoned into Jensen Huang's live interview at the All-In Summit and told several thousand executives on speakerphone that the robots will not be taking over — the second time in twelve hours he called AI risk a hoax.
- Axios lays out why a swift pause is unlikely: METR, floated by Anthropic as a third-party evaluator, has financial and personal ties to the labs it would audit, and Anthropic is expected to file this week to raise $100 billion at a $2 trillion valuation.
- Stuart Russell argues safety requirements beat schedule changes, because a pause with no specified endpoint commits nobody to meeting a concrete goal.
- Philipp Schmid of Google DeepMind says agent harnesses should shrink as models improve — two markdown files and a mounted cloud sandbox in place of Python orchestration, with Cursor cited as replacing ~12,000 lines of TypeScript with ~200 lines of agent files.

- Vercel's data-science agent doubled its eval scores after the team threw out prescriptive tooling and gave the model bash plus file read/write inside a sandbox.
- PostHog's counterweight: giving a model command execution is a "malware starter pack", so detection and blocking stay deterministic and fail-closed, with the model used only as a triage adviser.
- SemiAnalysis measured Vera Rubin NVL72 at up to 7x Blackwell's tokens per megawatt, above Huang's own 3x claim — while the organizer who got Monterey Park to ban datacenters is now running for city council.
- Christine Lagarde sets a deliberately modest bar for European AI: models good enough for most tasks, running domestically, so being cut off stops working as leverage.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Yesterday afternoon in Los Angeles, Jensen Huang was on stage at the All-In Summit, in the middle of a live interview, when he took a phone call from President Trump and held it up on speakerphone. Several thousand executives, investors and founders heard the president tell them that fears of an AI takeover are a hoax, and that the robots will not be taking over. NBC's Hallie

Jackson reported it for TODAY, Al Jazeera ran the clip, and Zachary Basu at Axios has the fullest account of what was happening around it.

- axios.com
- axios.com
- today.com
- aljazeera.com
- theguardian.com
- x.com
- youtube.com
- axios.com
- x.com
- techmeme.com
- theguardian.com
- techmeme.com
- x.com
- x.com
- x.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- techmeme.com
- techmeme.com
- techmeme.com
- theguardian.com
- theguardian.com
- techmeme.com
- techmeme.com
- techmeme.com
- techcrunch.com
- x.com
- courtlistener.com

2. Damra Vol 00:00:35

I keep coming back to the staging. He didn't call into a press conference. He called into Huang's interview, and Huang put him on the speaker, so the president is addressing a room full of the

people building this, with no moderator and no follow-up question. I've watched plenty of tech-conference theater and I can't remember another version of that one.

3. Lenar Kess 00:00:55

   We've got six things today, and that's the first, because it's the political answer to what happened over the weekend. Then the harder question sitting underneath it, which is who's actually qualified to grade a slowdown. Then a run of talks from the AI Engineer conference where three different companies made the same architectural argument. After that, Nvidia's Vera Rubin efficiency numbers next to a city council race in Monterey Park. Christine Lagarde on European models. And a short one at the end about Google engineers getting access to Claude.

4. Lenar Kess 00:01:27

   Quick setup, one line, because we spent Sunday on it. Dario Amodei published a thirty-thousand-word essay called We Must Pace the Frontier, arguing for a deliberate slowdown at the frontier, and Sam Altman, Elon Musk and others came out broadly in the same direction. That was the weekend. Then Monday morning, Trump posted on Truth Social that the only control AI needs is, quote, a STRONG AND SMART — high I.Q. — PRESIDENT, and the U.S.A. has that, in spades. Then, right after that: WHOEVER WINS AI, WINS. We are leading China, and all others, and will continue to do so.

5. Damra Vol 00:02:06

   All caps in the original, which you can hear in the rhythm of it. And it's aimed at Amodei personally. Somewhere in the same run of posts he takes a swipe at him for, quote, now pretending to be a perfect little angel. That reads less like a disagreement about regulation and more like something he's been carrying around for a while.

6. Lenar Kess 00:02:25

   Then Monday afternoon, twelve hours later, the second one. He called fears of AI taking over the world and destroying humanity a HOAX — his capitals — and put that in the same bucket as the Russia investigation and his own impeachments. Which is the same catalogue that already holds the Epstein files and the affordability crisis. He added that under his presidency, AI development won't be stopped by what he called brilliantly run Destructive Forces.

7. Damra Vol 00:02:52

   This isn't a fresh quarrel either. Axios notes the administration has clashed with Amodei and Anthropic repeatedly. The Pentagon designated the company a supply-chain risk, and stalled the release of its Mythos model. So when the president writes that we already have tremendous criminal

and regulatory power over these companies, that isn't abstract. There's a track record of using procurement as the lever, and everyone in that room yesterday knows it.

8. Lenar Kess 00:03:18

   That's the exact line, by the way: We already have tremendous CRIMINAL and REGULATORY power over these companies. And then one that needs a caveat attached. He said his administration has stopped AI, quote, people, from doing bad, or potentially bad, quote, things. His quotation marks, not mine. He didn't elaborate. Axios asked the White House about it Monday and got no response. So it sits there unexplained, and I'm not going to treat it as a fact about anything that actually happened. He also said there's a SICK conspiracy going on against AI and Data Centers, and the only one who's happy about it is China.

9. Damra Vol 00:03:56

   That conspiracy version has been circulating on his side of the valley for a while. David Sacks — Trump's AI adviser, and one of the All-In hosts, which is a tidy little loop — has described AI alarmism as the work of a well-funded Doomer Industrial Complex that includes former Biden officials. When the warnings went up last week, he and his co-hosts suggested the whole backlash was a coordinated psyop tied to election season and to the rise of open-source models.

10. Lenar Kess 00:04:24

    Meanwhile Congress is moving the other direction. The Guardian reports a rare bipartisan backlash, with Democrats and some Republicans pushing for rules on the largest AI companies. And the reason concerns hit this pitch is that the tech leaders raised the alarm themselves. Ted Lieu was posting about it Monday night. So we've got the unusual situation where an industry is asking to be regulated and the president is telling that industry it's imagining things.

11. Damra Vol 00:04:52

    Which inverts the normal script badly enough that I don't think either side has a plan for it. Regulation usually arrives because an industry resists it, and the resistance is what gives legislators something to push against. Here the labs wrote the argument themselves. A voluntary pledge among four companies still needs an enforcement environment — an evaluator, a reporting requirement, something with teeth — and the people who'd have to build that just called the premise fake on a speakerphone.

12. Lenar Kess 00:05:21

    Here's the fact that complicates the tidy version of this. China's Foreign Ministry also criticized calls to slow AI development on Monday. Guo Jiakun, a ministry spokesperson, told reporters — this is via CNBC — quote, Fear-mongering, confrontation, competition will just disrupt the process of

global AI governance. So Washington and Beijing arrived at opposite-sounding objections to the same proposal on the same day.

13. Damra Vol 00:05:48

    And I don't think that's a coincidence. Neither government wants the pacing decision made by four companies in a room in California. Trump's version is that the decision belongs to him. The ministry's version is that it belongs to a multilateral process it would rather run. Both of them are declining the same offer, for reasons that have nothing to do with the danger itself.

14. Lenar Kess 00:06:09

    Second thing, and this is the one that decides whether anything happens. Axios published a piece this morning with a blunt title — why a swift AI pause is unlikely: no one trusts AI companies. The argument isn't that people disagree about the danger. It's that almost nobody trusts the parties who would be doing the measuring, and that includes the safety organizations, because most of them came out of the labs.

15. Damra Vol 00:06:33

    The concrete version of that is METR. Anthropic has pointed at METR as a possible third-party evaluator — it's the group that evaluated the Hugging Face incident at OpenAI. And competitors point at a list of financial and personal ties between METR and the labs it would be auditing, and call that a non-starter. Susan Zhang flagged on X that METR disclosed a conflict of interest inside a three-hundred-page report. Technically disclosed. In practice you have to go looking for it.

16. Lenar Kess 00:07:07

    Gavin Baker put the objection at a higher level on X on Sunday. His line was that he doesn't want a few humans in control of intelligence. He also said Amodei was surely sincere in the essay, and that the moves would be good for his business over the long term. Both of those at once, and I don't think those two things contradict each other.

17. Damra Vol 00:07:27

    The money is where people keep tripping. As soon as this week, Anthropic is expected to file for what would be the largest initial public offering in history — raising a hundred billion dollars at a two trillion dollar valuation. Axios has the best sentence anyone's written about that duality: picture Oppenheimer meeting with sell-side analysts. A company whose own employees believe the product might destroy humanity is about to go sell it to the public markets.

18. Lenar Kess 00:07:55

And the pause talk moved money already. Global markets dipped Monday as investors worked out what a real slowdown in the one reliable growth engine of the year would do to everything downstream. Which is its own answer to why self-regulation is difficult here. Axios reaches for Meta's Oversight Board as the precedent and calls it largely toothless. That's the model currently on the table.

19. Damra Vol 00:08:18

    One detail in that piece I had to read twice. Axios names the four labs that agreed as Anthropic, OpenAI, SpaceX and Google. SpaceX. Not xAI — <emphasis>SpaceX</emphasis>. Either that's a slip in the copy, or the corporate structure over there has moved again and I missed it. And if you're designing an evaluator regime, which legal entity actually signed matters a great deal, because that's the entity you'd be writing obligations against.

20. Lenar Kess 00:08:46

    Matt Levine at Bloomberg came at it from the antitrust side. His point is that a coordinated slowdown among the leading producers looks, from a certain angle, like an agreement to limit output — and limiting output preserves margins on frontier models. He isn't accusing anyone of running a cartel. He's noting that what safety people want and what a cartel wants can be described in identical language, and that's a serious problem for whoever has to write the actual rule.

21. Damra Vol 00:09:14

    There's a third answer circulating that isn't either of those. Blake Scholl argued on X that you don't need a new regime at all, because existing liability law already covers the harms, and Section 230 is the distortion to fix. Yishan came at the same question from the direction of regulatory failure. Neither of those is a safety argument exactly. They're both saying the machinery exists already and nobody is using it.

22. Lenar Kess 00:09:40

    Stuart Russell has the sharpest counter, in the Guardian, and the headline is the argument: AI safety requires more than just slowing our pace. Requirements are non-negotiable, he writes, and they depend on meeting concrete goals rather than adjusting a timeline. He opens by pointing at Jacob Coxon's resignation from Anthropic as what set this week off. And he's right about the mechanics — a pause with no specified endpoint commits nobody to anything measurable. It's a schedule change.

23. Damra Vol 00:10:08

    Separately — and I do mean separately, these are two facts that happen to be adjacent on the same day — Bilal Chughtai resigned publicly from Google DeepMind's AI Safety and Alignment team.

Debby Wu has it at Bloomberg. His line was, quote, I earnestly believe that AI has the potential to kill us all. He said humanity is running out of time to prevent it.

24. Lenar Kess 00:10:32

    Last thing in this one, with a caveat bolted on. Kevin Bass has been calling on X for a congressional investigation into Anthropic's finances. That's a self-published thread. It isn't a filing or a referral, and no committee has said a word about it. I'm mentioning it because it's part of the noise any evaluator proposal now has to survive, not because there's a document sitting behind it.

25. Lenar Kess 00:10:55

    The next one is a different register entirely. A batch of talks from the AI Engineer conference went up, and three of them, from three different companies, make the same architectural claim. Philipp Schmid at Google DeepMind puts it most directly: as models get better, the harness around them should be contracting, not expanding. If adding a new capability makes your codebase more complicated, something is wrong with the architecture.

26. Damra Vol 00:11:20

    He traces the whole arc, which is useful if you lived through it. You started with hand-written Python loops, explicit schema definitions, function routing and state management. Then frameworks abstracted away the boilerplate and generated schemas off your function signatures, but you were still writing custom tool implementations for everything. His version now is two markdown files. One holds the system instructions and rules, the other holds the skills and context. Then you mount the sources you care about — a GitHub repository, say, or a storage bucket — into an isolated cloud Linux sandbox.

27. Lenar Kess 00:11:57

    There's a network proxy that injects credentials on request, so the model never holds them, and you configure the domain restrictions yourself. On top of that sits an Agents API. You get a reusable agent identifier with its configuration attached, the session state lives on the server, and context compaction happens automatically across turns. And there's an Interactions API that throws out turn-based conversation history in favor of a step-based timeline, where reasoning and function calls and results are first-class steps instead of everything getting crammed into message roles.

28. Damra Vol 00:12:31

    The numbers he cites are the ones people will repeat. Cursor replaced roughly twelve thousand lines of TypeScript orchestration with about two hundred lines of agent files. LangChain has rebuilt its Deep Research system three times a year because of complexity. One company cut eighty percent of

its custom tools and got faster, more accurate responses. I'd hold all of those loosely — these are conference-stage claims from people with something to sell, not published benchmarks.

29. Lenar Kess 00:13:01

    Andrew Qu, Chief of Software at Vercel, tells the same story from inside a product. Their data-science agent, D0, started as a single mega-prompt for generating SQL. That wasn't enough. So they went to a multi-agent chain, with a planner, an executor and a reporter. Then they collapsed that into one stateful agent managing its own memory across roughly a hundred steps. Evaluation scores looked strong.

30. Damra Vol 00:13:28

    And then they handed it to real beta users and it fell over. His word for the cause is prescriptive — the tooling told the model exactly how to do each thing, and real questions didn't match the script. So they went and studied Claude Code and Opus 4.5, noticed those systems expose almost nothing, just bash and file read and file write inside a sandbox, and rebuilt D0 that way. They dumped the semantic layer into the sandbox as files and let the agent go read it. That doubled their evaluation scores.

31. Lenar Kess 00:14:00

    They also added a job that distills recent queries into about a hundred reusable skills, so the agent starts each run with context nobody mapped by hand. And they open-sourced the result as Eve. It's an agent framework with Next.js-style conventions, so you define an agent by making directories for skills, tools and channels. Underneath, it runs on Vercel's Workflows for state, their Sandbox for execution, and Connect for short-lived OAuth tokens. A beta customer, Aura, rebuilt a service-testing agent on it and reported fewer execution steps and a higher success rate than their off-the-shelf Claude Code deployment.

32. Damra Vol 00:14:38

    Now the counterweight, and I'm glad it was on the same program. Sarah is a context engineer at PostHog. She built The Wizard, a command-line agent that reads your codebase, installs the right SDK, instruments your events and configures dashboards in five or six minutes. Around eight thousand developers use it every week. Her description of what she'd actually shipped is the line of the day: giving a large language model command execution is a malware starter pack.

33. Lenar Kess 00:15:06

    So she audited it. The original defense was prompt steering plus a tightly bounded allow list. Bash was denied by default, installations were restricted to vetted packages, environment variable access was blocked, and secrets got routed through a vault. Reasonable as far as it goes. The threat she

found underneath was the supply chain: poisoned documentation or code comments sitting in open-source repositories, ingested by her own context engine, and delivered as prompt-injection payloads to eight thousand people at once.

34. Damra Vol 00:15:38

    Her answer is where I'd send builders. She wrote a standalone scanner called Warlock that uses YARA rules — deterministic pattern-matching rules, each with test suites for both matches and non-matches — to detect malicious content in sources and in runtime inputs, without executing anything. It catches sub-agent bypasses and accidental leakage of personal data. It also throws a lot of false positives off benign demo code. So she added a model on top as a triage adviser and nothing more, reviewing findings that were already unblocked. Detection and blocking stay deterministic and fail closed. The model never gets to be the gate.

35. Lenar Kess 00:16:18

    One more from the same program, because the numbers are unusually concrete. Caitlyn and Angela from Anthropic argue that tokens aren't fungible and should be assigned jobs. They name four. The executor does the task, the adviser gives real-time guidance, the grader scores output against a rubric, and the dreamer reads execution transcripts afterward and writes learnings into memory for the next round. On a financial-analysis benchmark, a one-shot executor used thirty-nine thousand tokens and got fifteen percent accuracy. Hold the budget fixed at around six hundred thousand tokens and that same executor reaches seventy-six percent. Advising and grading get to roughly eighty-nine or ninety.

36. Damra Vol 00:17:01

    The pass-fail number is the one I'd write down. Score it the way a bank would, where the answer has to be entirely right or it's worthless, and the plain executor succeeds in about forty-two percent of runs. Getting one perfect answer takes about three runs on average, so call it one point eight million tokens. That's the actual price of skipping the adviser.

37. Lenar Kess 00:17:24

    What else came out of that program?

38. Damra Vol 00:17:26

    Two more short ones. Mike Chambers at AWS warns against what he calls slop ops, meaning letting agents provision cloud resources straight from prompts instead of generating infrastructure-as-code you own. And there's a talk on Restate, which journals agent execution so a crash resumes at the exact point instead of starting the whole run over.

39. Lenar Kess 00:17:47

    Let's do compute economics. Bryan Shan at SemiAnalysis ran inference tests on Nvidia's Vera Rubin NVL72 and measured up to seven times better token throughput per megawatt than Blackwell, on a 1.6 trillion parameter DeepSeek model. Jensen Huang's own public claim was three times, for models in the one to three trillion parameter range.

40. Damra Vol 00:18:10

    A vendor undershooting its own benchmark by a factor of two is unusual enough to stop on. The SemiAnalysis headline actually reads Jensen sandbagging performance again, so they think it's a habit rather than a one-off. And I'd take the measurement seriously in one narrow sense: tokens per megawatt is the number that matters now, because power is the input you can't just order more of when you need it.

41. Lenar Kess 00:18:35

    Caveat on that — SemiAnalysis is paywalled and I've only got the headline figures, so seven and three are all I'm going to quote. Around the same time, a Dutch inference-chip startup called Euclyd raised a Series A of more than two hundred million euros, about two hundred thirty-one million dollars. Samsung co-led it, alongside Somerset Capital, the Scaleup Europe Fund and Innovation Industries. Kai Nicol-Schwarz has that at CNBC.

42. Damra Vol 00:19:02

    Samsung backing a European alternative to Nvidia's GPUs is the interesting half of that one. And SemiAnalysis published a separate deep dive the same day on robot inference. It covers on-device versus data center, how efficient the silicon and memory are, how Jetson Thor compares against a B300 on total cost of ownership, and something they call the network wall. If you're trying to work out where embodied inference has to run, that's the primer to read.

43. Lenar Kess 00:19:31

    Now the other end of the same conversation. The Guardian reports that anti-datacenter organizers around the country are converting community anger into local election campaigns. Monterey Park, just east of Los Angeles, became the first US city to ban datacenter construction, and that took four months of teach-ins and canvassing. Steven Kung, who helped organize it, is now running for city council against the incumbent who supported the development.

44. Damra Vol 00:19:58

    He said two things to the Guardian. The datacenter coming to my neighborhood turned my life upside down. And: I saw that there's a dismantling of democracy at a local level. We went deep on

siting and power costs yesterday so I won't do that again. The new element is that the fight now produces candidates. You oppose a facility, you win or you lose, and either way you come out of it with a list of neighbors who'll knock doors for you.

45. Lenar Kess 00:20:25

    Christine Lagarde, president of the European Central Bank, told the Guardian that Europe has to develop its own AI models and build more datacenters, so that dependency on American or Chinese providers can't become negotiating leverage against it. Her phrase for the requirement was models good enough to carry out most tasks, running on domestic infrastructure. If Europe has that, she argues, the threat of being cut off loses its force.

46. Damra Vol 00:20:51

    The bar is what's new there. Almost every European sovereignty argument I've heard for two years has been about frontier parity — catch up or lose. Hers isn't. Good enough for most tasks is a far lower target, and it's reachable by fine-tuning open weights on European hardware, which is a completely different industrial program from building a frontier lab. It's also a central banker's version of the argument. She's reasoning about leverage in a negotiation, not about capability.

47. Lenar Kess 00:21:21

    Separate story, same axis. The Financial Times reports that China has tightened overseas travel controls for its own citizens, through exit bans and fines under a sweeping new law, as Xi moves to secure state secrets, advanced technology and highly skilled workers. I'm putting it next to Lagarde rather than inside her argument, because they aren't the same move. One state is trying not to depend on anybody. The other is trying to keep what it already has from walking out the door.

48. Damra Vol 00:21:51

    And one number while we're in the neighborhood. The Information reports ByteDance's first-half net profit fell by a single-digit percentage, to twenty billion dollars. Revenue over the same stretch rose about thirty percent, to a hundred and twenty billion. Full-year 2025 revenue was up twenty-nine percent, to roughly two hundred billion. Revenue climbing while profit slips is what it looks like when you're buying compute at that scale and haven't finished turning it into products yet.

49. Lenar Kess 00:22:21

    Two short ones and then I'll let you go. Hugh Langley at Business Insider reports Google is giving all of its engineers access to Anthropic's Claude Opus 5 through Antigravity. Google's statement alongside it: Gemini remains our primary and foundational model for internal development.

50. Damra Vol 00:22:39

Which is a sentence written by a communications team and an engineering organization in the same breath, and you can see the seam in it. But I'd read the procurement decision rather than the press line. Somebody at Google measured how much slower their engineers were without Claude and decided that number was larger than the embarrassment of admitting it.

51. Lenar Kess 00:22:58

    There's a counterpoint going around that Nvidia and Palantir are limiting internal use of Anthropic models. I'll tell you exactly what that is: a one-line post from the Watcher.Guru account with no linked filing, statement, or named source. I'm mentioning it so you've heard it and so you know there's nothing under it yet.

52. Damra Vol 00:23:17

    The other one is Salesforce and Nvidia shipping Koa, a reasoning model built on Nvidia's open-weight Nemotron and trained for sales, marketing and customer-support work. TechCrunch's headline calls it everything the AI labs should fear, which I'm not adopting. A vertically trained model on open weights is a procurement option, and it'll be judged on whether it closes tickets faster than the general-purpose one.

53. Lenar Kess 00:23:42

    Last item, and it's a loose end. Six new entries posted yesterday in the consolidated OpenAI copyright litigation in the Southern District of New York — docket numbers 1957 through 1962. On the public docket all six read Original document, with no summary attached to any of them. Six filings in a single day on a consolidated case is probably something. I can't tell you what, and I'm not going to guess at it.

54. Damra Vol 00:24:10

    Somebody should read them before tomorrow. Six entries at once usually means either a scheduling order with a batch of exhibits attached, or a dispositive motion where everything supporting it came in together. Those are very different events for the case. What we have is six numbers and a placeholder, which is nothing.

55. Lenar Kess 00:24:28

    Somebody will, and if the documents say anything, we'll have it tomorrow. The evaluator question is the one still sitting open. Four labs agreed on a direction over the weekend, and there's still nobody that all four of them, Washington, and the people who'd have to write the rule would accept holding the stopwatch. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic