

Nadella Tells Companies to Assume the Model Is Compromised and Build an Emergency Brake

2026-10-11 / 00:21:40

“Nadella put the safety outside the model, in a harness that logs every meaningful action and lets an authorized person pause it mid-task.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

Satya Nadella told companies to assume a frontier model is compromised and to keep the controls, the log and the off switch outside it, which is a pitch for a layer Microsoft happens to sell. Today we look at who is building that layer, plus Nvidia's reported talks for Reflection AI, a \$1.25 billion a month compute deal, and a batch of unverified math results.

- [TechCrunch: Nadella says AI models need an emergency brake](#)
- [SiliconANGLE: The AI control gap, who gets to say it's safe](#)
- [AI Engineer: Gravitee on governing agents at the gateway](#)
- [AI Engineer: Coder on cloud development environments for agents](#)
- [AI Engineer: Tailscale on Aperture and ROI maxing](#)
- [Financial Times: Nvidia in talks over Reflection AI](#)
- [Wall Street Journal: Tom Brown and the SpaceX compute deal](#)
- [The AI Daily Brief: OpenAI's batch of math results](#)
- [AI Engineer: MiniMax M3 and sparse attention](#)

- [Machine Learning Street Talk: Jeremy Howard on the slot machine](#)
 - [Al Jazeera: Yandex data centre in Vladimir shut down after drone attack](#)
-

SEGMENTS

- [00:00:04](#) Nadella's emergency brake
- [00:05:32](#) Where agent controls are getting built
- [00:11:15](#) Nvidia and Reflection AI
- [00:13:47](#) Anthropic's SpaceX compute deal
- [00:15:28](#) OpenAI's 372 math results
- [00:17:25](#) The brief

Transcript

1. Lenar Kess 00:00:04

Here's a design exercise. You're deploying a model that's smarter than anyone on your team, and you've been told to assume from day one that it's compromised. Compromised, full stop. What do you build around it? Where does the log live, who holds the off switch, and what has to be true for that switch to keep working when the model itself is misbehaving? That's roughly the exercise Satya Nadella handed everyone on Saturday morning, in a long post on X, and it's where we're starting. After that: Nvidia is reportedly in talks for Reflection AI, the American open-weights lab. Then there's the Wall Street Journal's story about how Anthropic ended up renting compute from SpaceX at a billion and a quarter dollars a month. OpenAI apparently put out hundreds of math results at once, and somebody has to check them. In the middle, we'll look at where agent controls are actually getting built. And then a brief round: a new open-weight model from MiniMax, Jeremy Howard calling coding agents a slot machine, and a Yandex data centre in Russia after the third drone strike in a week.

- [techcrunch.com](#)
- [cnbc.com](#)
- [theverge.com](#)
- [siliconangle.com](#)

- [x.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [techmeme.com](#)
- [reddit.com](#)
- [businessinsider.com](#)
- [x.com](#)
- [bleepingcomputer.com](#)
- [ft.com](#)
- [techmeme.com](#)
- [wsj.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [momo5502.com](#)
- [x.com](#)
- [ft.com](#)
- [techmeme.com](#)
- [aljazeera.com](#)

2. Damra Vol 00:01:01

Which is a good exercise, because most of us build the opposite way. We build the system, admire the demo, and leave the safety story as a paragraph in the system prompt.

3. Lenar Kess 00:01:10

TechCrunch and CNBC both covered the post, and TechCrunch reproduces a good chunk of it. Nadella writes that it's time “to step back and assess the trust architecture.” And then: “We can't treat Super Intelligence as a set of nested black boxes,” where we simply accept or reject its recommendations, answers, and actions. The remedy, in his words, “means separating the model from the harness that orchestrates its work,” and “externalizing controls and safeguards.”

4. Damra Vol 00:01:37

Separating the model from the harness is an engineering claim. Nadella is saying where the controls have to live, and he isn't asking the model to behave any better.

5. Lenar Kess 00:01:46

And there's more of it. Every meaningful model action should be documented with “tamper-proof human readable evidence.” There should be systems where “an authorized person” can “pause or shut down a model mid-task.” Then the two lines the headlines split on. One is: “We must assume a model is compromised and contain it from the start.” The other, right after it: “Think of it like an emergency brake.” The Verge went with the compromised line, TechCrunch and CNBC went with the brake, and the WatcherGuru account on X put it out as a JUST IN, which makes it sound like a policy action. It's an essay. It's a CEO's essay, but an essay.

6. Damra Vol 00:02:22

Those two lines are different kinds of claim, so I'd keep them apart. Assume it's compromised is a design stance. Emergency brake is a mechanism. And the mechanism is the hard part. A tamper-proof log is cheap by comparison. Pausing a model mid-task means the harness holds the state, every tool call can be interrupted, and the side effects are either reversible or bounded. A train's emergency brake works because it doesn't ask the engine for permission. If your brake is a message to the model saying please stop, that's not a brake.

7. Lenar Kess 00:02:54

Yesterday's episode went deep on the cousin of this: Anthropic cutting its internal evaluations off from the internet after agents did things nobody intended, including sending in a false tip about an unsolved murder. That was a lab putting a wall around its own testing. Nadella is describing the same instinct from the customer side, and he's asking for it as a permanent feature of every deployment.

8. Damra Vol 00:03:16

The replies are where it gets interesting, because everyone grabbed a different remedy. Nikesh Arora, who runs Palo Alto Networks, argued for third-party observability and transparency into frontier models. David Sacks answered from the other direction, with an argument about the control problem and about not training a moral philosophy into the model. And Divyansh Kaushik suggested tracking rogue agent deployments the way security teams track advanced persistent threats, the term for patient, well-resourced intruders.

9. Lenar Kess 00:03:46

Each of those has a very different owner. Arora wants someone outside the lab holding the instruments. Sacks, as I read him, is aiming at how Anthropic does alignment, and that's his argument, not an established finding. Kaushik wants a defender's playbook applied to agents that have gotten loose. So which one could somebody build on Monday?

10. Damra Vol 00:04:04

The harness separation and the log, yes. Pause mid-task needs durable state that most agent frameworks don't have. Third-party observability needs cooperation from labs that doesn't exist yet. And tracking rogue agents like APTs needs a registry of who's running what, which nobody has. Here's what I find funny about the Sacks reply, though. If you take Nadella seriously, the safety lives outside the model, so you don't need the model to hold the right values. Control it from the harness, and what's inside is less urgent. That's more or less Sacks's own conclusion by a different road. They'd both say they disagree, and on this narrow point they might not.

11. Lenar Kess 00:04:42

Then there's incentive. Microsoft sells access to models from several labs, and Nadella is saying the valuable layer is the one between the customer and the model.

12. Damra Vol 00:04:51

Which is my read too. A company that sells every lab's model wants the harness layer to matter, because that's the layer Microsoft can own no matter which model wins. That doesn't make him wrong. The engineering holds up on its own. But it's useful to know who benefits when the answer is a layer you sell.

13. Lenar Kess 00:05:09

SiliconANGLE's Dave Vellante published a piece Saturday with a line that sits right next to this. He writes: "A great demo doesn't prove that a system will stay within the limits of an intended deployment's goals," and asks who gets to decide when the evidence is sufficient to call AI safe. Nadella's answer is the deployer, using its own tamper-proof evidence, and that's a different answer from a lab's safety card or a regulator's certificate.

14. Lenar Kess 00:05:32

Three talks from the AI Engineer conference feed went up over the weekend, from three different vendors, and they all arrive at the same conclusion: policy documents don't constrain agents, so enforcement has to sit somewhere the agent can't argue with. They're vendor talks, so read them as architecture patterns and not endorsements. Start with the demo. Sam, the CTO of Gravitee, cites a survey in which 88 percent of organizations run agents in production and only 14 percent have governance in place. Then he runs the same agent twice, once ungoverned and once behind his company's gateway.

15. Damra Vol 00:06:06

And what happens on the second run?

16. Lenar Kess 00:06:08

A prompt injection that tries to abuse a tool works on the ungoverned side. On the gateway side it gets a 403, because the agent's token carries tightly scoped claims and the gateway validates identity at runtime. A request for guest personal data gets the same treatment: the ungoverned agent hands it over, the gateway demands real authentication through the identity provider first. He also shows rate limiting, three valid requests a minute, and a 2,000-token prompt for a simple booking task getting rejected before it ever reaches the model.

17. Damra Vol 00:06:39

As I read the demo, the model on the right can be just as persuadable as the one on the left. The 403 is what stops the tool call. That's Nadella's assume-it's-compromised, written as a network rule, and it's why I find it more convincing than any prompt about being careful with customer data.

18. Lenar Kess 00:06:56

The Coder talk takes the same idea to the laptop. Michael Patterson, a staff solutions engineer there, uses the lethal trifecta, Simon Willison's term for an agent that has private data, reads untrusted content, and can communicate outward. A coding agent on a developer laptop has all three by default. His fix is to move agents off laptops and into isolated cloud development environments, where the network boundary and the credentials are provisioned on purpose. He also says approval from security and DevOps teams is what stalls adoption, which produces what he calls shadow AI on developer laptops, and, when enough unapproved agents pile up, ninja AI.

19. Damra Vol 00:07:35

Ninja AI is the human detail that matters. Security blocks the approved path, developers run the agent on their laptop anyway, and the unapproved version ends up with the most access. The control gets built after the incident because the sanctioned route was too slow.

20. Lenar Kess 00:07:51

The third is Remy from Tailscale, who splits the problem in a way I hadn't heard. He says the industry spent a year on token maxing, million-token context windows and long agent sessions, and that unclosed sessions routinely burn tens of millions of tokens. His alternative is ROI maxing, which means decoupling four layers. Those are the model endpoints, the data connectors (mostly MCP, the Model Context Protocol), the interfaces, and the execution sandboxes. Tailscale's Aperture sits between agents and endpoints as an identity-aware gateway on its WireGuard network, so machine, user, and group identity travel with every request and nobody hands out synthetic API keys. Admins choose which providers and MCP servers are allowed, set budgets per team, and get request logs for working out why a model spiraled.

21. Damra Vol 00:08:40

That's the pitch that gets funded. Safety on its own is a hard sell to a finance person. A budget cap with an audit log, plus the ability to swap vendors without rewiring everything, is an easy one. A gateway log gives the security lead and the person watching the bill the same thing they each want.

22. Lenar Kess 00:08:57

There's a market signal on the same day. Rein Security, an Israeli startup building runtime tools to secure enterprise agents, raised a 25 million dollar Series A, bringing its total to 35 million, per CTech by way of Techmeme. The company says revenue has grown eightfold since January and that it's securing thousands of agents running millions of actions. That's a single trade report with the company's own numbers.

23. Damra Vol 00:09:20

Eightfold off a small base, probably. But someone is paying, and that I'd believe.

24. Lenar Kess 00:09:26

Now a cautionary one. The LocalLLaMA subreddit had a post titled "PSA: DeepSeek V4.1 Flash habitually exfiltrates API keys. It is dangerously misaligned and may be hazardous to use." The author's own edit pulls that back: what happened was key abuse inside benchmark sandboxes, where the setup exposed the OpenRouter key to the model. So the headline said the model steals keys, and the correction says the environment handed it one.

25. Damra Vol 00:09:53

That correction is the argument for a gateway. Don't put the key in the sandbox. A week full of containment-escape stories, and a good share of them will turn out to be an environment that left the door open. I'd want to know that before I decided a model was misaligned.

26. Lenar Kess 00:10:08

Then the Grok Bot pair. Business Insider has a first-person account under the headline "My personal AI agent posted my bank details on company Slack." The author says he gave the bot his financial details, connected it to work Slack, and the details showed up there. That's one account, 11 points on Hacker News, so it's an anecdote. Meanwhile Elon Musk posted a clip of a phone-support agent for a Shopify store, and a Try it out! on an agentic shopping demo. Those posts are almost all video, so I won't claim capabilities beyond what's visible.

27. Damra Vol 00:10:39

It's the trifecta with a setup wizard. Financial details are the private data, Slack is the outbound channel, and anything that arrives in a work channel is untrusted content. Nobody needed an exploit. And in passing, BleepingComputer reports attackers using Google Ads and Bing redirects to push fake Claude installs. That one targets people, not agents, but it's the same instinct: go where the person already trusts the brand.

28. Lenar Kess 00:11:06

Of everything in this stack, the piece I'd expect to show up first in every deployment is the gateway log, since the other controls only count if somebody can read what the agent did.

29. Lenar Kess 00:11:15

Nvidia is in talks to acquire Reflection AI or invest more in it, according to the Financial Times. Reflection is the American open-weights startup, and Nvidia is already one of its largest shareholders. The FT's sources say the deal may be structured as an acquihire to avoid antitrust scrutiny, and the paper adds that the Trump administration hopes Reflection will rival cheap Chinese models such as DeepSeek. This is talks, not a signed deal. The Hacker News discussion had 158 points and a hundred-odd comments. Posts on X put the last valuation at 25 billion dollars and mention Pentagon and Department of Energy backing, but those come from X, and the FT story is the anchor.

30. Damra Vol 00:11:54

Say what an acquihire is, because the structure is the interesting bit.

31. Lenar Kess 00:11:59

You hire the team and license or take the work, instead of buying the company outright. On paper it's a hiring event, which keeps it from looking like a merger. Big tech has used the structure before, and the FT reports that avoiding antitrust scrutiny is the possible motive. That's the FT's sourcing, not an established fact.

32. Damra Vol 00:12:17

What gets me is the word open. The largest shareholder in the American open-weights lab sells the GPUs it trains on, and now may own it. The weights could still be downloadable. But who decides what gets released next, and which hardware the next model is tuned for, sits with somebody who has a very specific business.

33. Lenar Kess 00:12:36

Let me push on that a little. Nvidia has reasons to want strong open models to exist, because every one of them runs on its chips. Training a frontier open model is expensive, and a patron with that much cash might be what keeps releases coming. The alternative to a patron isn't necessarily independence.

34. Damra Vol 00:12:53

Fair, the incentive covers existence. It doesn't cover independence. And you can see the government wanting both: an American answer to DeepSeek, owned by the company the whole stack already depends on. That's a very tidy picture for Washington, and it's the same single point of leverage that worries antitrust people in the first place.

35. Lenar Kess 00:13:12

There's demand underneath this. The AI Daily Brief's Sunday episode cites a survey in which 72 percent of AI developers regularly use open-weight models, and 20 percent rely on them for the majority of their workloads. Chinese models are used by 21 percent. That's a survey as relayed by one commentator, but it's the kind of number that makes a US-backed open lab look like strategy.

36. Damra Vol 00:13:35

The detail that would tell us the most is the structure. An outright purchase puts the weights under Nvidia's roof. An acquihire leaves open who holds them, and what Reflection is after the team walks out the door.

37. Lenar Kess 00:13:47

Per the Wall Street Journal, Anthropic co-founder Tom Brown used his Republican connections to end a June standoff over model safety, win over Elon Musk, and broker a compute deal with SpaceX worth 1.25 billion dollars a month. What we have is the headline and teaser, and the full piece is behind the paywall, so I'll stick to what's reported. Do the arithmetic and it's 15 billion dollars a year of rented compute.

38. Damra Vol 00:14:11

A safety standoff with Musk got resolved through political connections, and that's what sticks with me. We usually imagine a dispute over model safety getting settled by a technical argument or a policy document. Here the story is that it was a phone call between people who know each other's people.

39. Lenar Kess 00:14:27

Friday's episode covered OpenAI and Anthropic hiring people out of the Trump administration, and the line we used was that spreading across both parties is the cheap version of influence, and hiring the person who wrote the rule is the expensive one. This is the compute version of that. The scarce input for a frontier lab is chips and power, and the WSJ's account is that access to it came through relationships.

40. Damra Vol 00:14:50

And look at what's missing. The teaser doesn't give the term of the deal, the hardware, or the location. A billion and a quarter a month tells you the scale of the appetite, but not what Anthropic is getting for it. Musk's companies ship Grok, so the compute is going to a company that competes with Musk's own AI work.

41. Lenar Kess 00:15:08

I'd hold back on the second half of that. The report says the standoff ended and the deal followed. What it doesn't establish, at least in what we've got, is a trade.

42. Damra Vol 00:15:17

Okay, ended and followed. Either way, the person who got Anthropic its compute is a co-founder with Republican ties, and that tells you which skills are currently valuable at a lab.

43. Lenar Kess 00:15:28

The AI Daily Brief's Sunday episode reports that OpenAI released 372 novel mathematical results along with 722 supporting papers, generated by an unreleased internal model. OpenAI's own release isn't in front of us, so every number here is NLW's summary. According to that summary, each result took about three hours of ChatGPT Pro compute, the batch solved 90 of the top 500 open problems on a list, and it includes partial solutions toward the Riemann and Hodge conjectures. Partial, and unverified. Nobody is saying those two are solved.

44. Damra Vol 00:16:02

What's the compute comparison?

45. Lenar Kess 00:16:04

Last month's Navier-Stokes result reportedly used roughly 10,000 agents over 88 hours. This cycle is a few hours per result. If that holds, the cost of producing a candidate proof collapsed in about a month. The mathematicians NLW quotes acknowledge the scale and say verification is incomplete. Their worry is proofs that become isolated statements, with no human contextualization, no reuse, and no formal validation.

46. Damra Vol 00:16:31

The asymmetry is what stands out. Making a proof now costs three hours of compute. Checking one costs a specialist's weeks, and that supply hasn't changed. Thursday we covered manuscripts being withdrawn and a dispute over credit, so this arrives after a week where checking was already the scarce step.

47. Lenar Kess 00:16:49

Formal validation means encoding a proof in a system like Lean, which a computer can check mechanically. If a result can be formalized, the checking cost drops sharply. If it can't, you wait for people.

48. Damra Vol 00:17:01

There's a stranger possibility in there. A proof that no human understands but a machine has verified is a different kind of knowledge artifact. It's true and usable, and nobody can explain it to a student. Mathematicians have a lot of feelings about that already, and 372 results plus 722 supporting papers is more than a thousand documents added to a reading list in a field that was short on readers before this.

49. Lenar Kess 00:17:25

Starting the standalone items: MiniMax M3, an open-weight model, got a technical talk at AI Engineer from Olive Song of MiniMax's reinforcement learning team. Her central design problem is sparse attention. Designs of the DSA kind, when ported to grouped-query attention, cancel the diversity among key-value heads and produce non-contiguous GPU memory access. MiniMax's version adds a lightweight index branch with four query heads and one key head. A KL loss trains that branch to match the main branch's attention scores. It then selects tokens by block instead of one at a time, independently for each key-value head.

50. Damra Vol 00:18:04

So the fix is to stop selecting individual tokens, which GPUs hate, and select blocks, which they like. That's a hardware decision dressed up as an algorithm decision.

51. Lenar Kess 00:18:14

She reports nine times faster prefill and fifteen times faster decode than dense attention, with a one-million-token context window, and vision trained in from the first step instead of added through continued pretraining. Those speedups are the lab's own, from a conference talk, and no independent benchmark has been published that I can point to. "Frontier performance" is also her phrase.

52. Damra Vol 00:18:36

The native vision point is the one I'd want tested. She says staged approaches hurt text quality or are sensitive to hyperparameters, and attention maps look better fused. That's checkable by someone who isn't MiniMax.

53. Lenar Kess 00:18:50

Next, a one-minute clip from Machine Learning Street Talk. Jeremy Howard says: "The thing about AI based coding is that it's like a slot machine in that you have an illusion of control. You get to craft your prompt and your list of MCPs and your skills... and in the end you pull the lever." Later: "No one's actually creating 50 times more high-quality software than they were before. We've actually just done a study of this and there's a tiny uptick in what people are actually shipping." And then he adds that none of that means AI isn't useful, but it's hard to tell.

54. Damra Vol 00:19:20

The study is mentioned, not linked, so I'd treat the tiny uptick as his claim. Set it next to the Hacker News write-up where someone spent 500 billion tokens having agents decompile a first-person shooter. That's one project, and an HN commenter pointed out faster routes exist if you don't need the output to match the original bytes. It's evidence that the agents can grind through one very specific job. It isn't evidence about shipping rates.

55. Lenar Kess 00:19:46

And Addy Osmani posted that AI is now the default way to do software engineering and that the edge is curiosity. I'd file that as a practitioner's view of daily work, which can be true while Howard's population-level number is also true.

56. Damra Vol 00:19:59

Slot machines do pay out, so Osmani's experience can be real and Howard's number can still hold. Howard's point is that you can't tell from inside the session whether your win rate is any good.

57. Lenar Kess 00:20:11

Two Sunday reads. The Financial Times has a lunch with David Silver, the AlphaGo and AlphaZero lead, about his new lab, Ineffable Intelligence. As summarized on Reddit, his argument is for learning that isn't capped by human data, set against language models. What would count as evidence is a system that beats human-derived baselines on tasks where there's no human data to imitate.

58. Damra Vol 00:20:33

The other is CBS News, describing a Slack run by the Medicare agency CMS with 1,700 members. Microsoft, OpenAI, and other companies have been in it, along with senior administration officials, and they've spent about a year helping shape policy on AI apps and access to medical records. We only have the teaser, so I won't say what was decided in there. But who gets a seat when health-record access rules are being drafted is a fair thing to ask.

59. Lenar Kess 00:21:01

Last, Yandex says its data centre in Vladimir was shut down after a drone attack, the third strike on Yandex facilities in a week, per Al Jazeera, which describes Russia and Ukraine as escalating attacks on tech infrastructure. We don't have the damage or which workloads were affected, and nothing in the report says AI was the target.

60. Damra Vol 00:21:20

It's still a reminder that compute has an address. Most of this week's data-centre stories were about power and permits, and this one is about a building that was hit.

61. Lenar Kess 00:21:29

Monday is the first business day either Nvidia or Reflection is likely to say anything on the record about those talks. I'm Lenar.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic