Leave Safety to Us

2026-09-16 / 00:23:08

> *“A delay you disclose after you've shipped is unfalsifiable from the outside.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Four people at the top of this industry answered the pacing argument this week, and none of them answered it the same way: self-regulation by product maker, a coordinated pause, cross-lab peer review, or plain tort liability. Underneath the disagreement, a 154-page misuse report, a copyright default about to flip in Australia, and a pile of papers arguing you should use fewer agents, not more.

- Jensen Huang at Dreamforce, via TechCrunch
- Dario Amodei's brake analogy, in the Guardian
- Zuckerberg on slowing down, at CNBC
- Musk on cross-lab safety testing
- David Krueger: the plan isn't enough
- Anthropic's 154-page misuse report, walked through
- M.G. v. Altman: the motion-to-dismiss schedule
- OpenAI valuation talks, $1.2T against $1.5T
- Australia's leaked text-and-data-mining exception

- Cloudflare on splitting crawler identity
- Bohan Li on where voice latency actually lives
- Decomposition Buys Integrity, Not Yield
- Mo' Models, Mo' Problems
- TAME: token attribution and masking

SEGMENTS

00:00:04	Dreamforce answers back
00:05:03	The 154-page misuse report
00:08:26	Splitting a harmful task into harmless questions
00:10:43	The docket answer, and a $300 billion gap
00:12:22	Australia considers flipping the default
00:14:26	Where voice latency lives
00:18:03	Fewer agents, and 200 thousand tokens on a laptop

Transcript

1. Lenar Kess 00:00:04

   Dreamforce is a customer-relationship-management sales conference. A keynote, a trade floor, and a concert at the end. This year it turned into the room where the pacing argument got answered by name. Dario Amodei went on stage and gave the Guardian's Danielle Abril the analogy he's been using all week — you don't wait for the brakes to fail before you decide cars need brakes. And then, over the next two days, three other people at the top of this industry answered him, and not one of them answered the same way. We start there. After that, Anthropic's 154-page misuse report and a Microsoft paper on splitting a harmful task into harmless pieces. Then a docket answer we owed you from yesterday, Australia considering flipping a copyright default, where voice latency lives, and a set of papers arguing you should use fewer agents rather than more.

- techcrunch.com
- theguardian.com
- cnbc.com
- x.com
- x.com
- cnbc.com
- theguardian.com
- x.com
- techmeme.com
- techmeme.com
- aljazeera.com
- youtube.com
- techmeme.com
- blog.cloudflare.com
- x.com
- arxiv.org
- courtlistener.com
- courtlistener.com
- courtlistener.com
- techmeme.com
- theguardian.com
- theguardian.com
- theguardian.com
- siliconangle.com
- youtube.com
- arxiv.org
- arxiv.org
- x.com
- arxiv.org
- arxiv.org
- arxiv.org
- techmeme.com
- techmeme.com
- nbcnews.com
- x.com

2. Damra Vol 00:00:56

   Jensen Huang's version came with a headline attached. Julie Bort at TechCrunch ran it under the line: we don't need AI regulation, leave safety to us. The claim underneath is narrower than the

headline, and I think it deserves to be stated at its actual size. He's saying an AI system isn't an alien mind. It's hardware and software, built by people who understand the parts, and so safety gets engineered per product, by whoever makes the product. A car company makes a car safe. A chip company makes a chip safe. Nothing in that chain requires a regulator who understands the whole thing.

3. Lenar Kess 00:01:32

   Being made of parts we understand doesn't get you a predictable composed system, though. [pause] Aircraft are made of parts we understand. We still have a Federal Aviation Administration, and it exists because things break in the composition and in the cockpit, not in any single component anybody shipped. Huang's argument works at the level of a chip and stops working at the level of a deployed agent that three vendors touched.

4. Damra Vol 00:01:56

   Mark Zuckerberg sided with Huang. CNBC had him arguing against slowing down, and then David Sacks relayed a detail I keep turning over. Meta held Muse back for months over safety concerns, and Zuckerberg presented that as ordinary day-to-day work — what a responsible company already does without being told. Which, fine. But a delay you disclose after you've shipped is unfalsifiable from the outside. Nobody outside the company can audit it, check what the concern was, or tell whether the delay ran three months or three days. That's the exact category of commitment Amodei's proposal exists to replace, and it's being offered as evidence the proposal is unnecessary.

5. Lenar Kess 00:02:38

   Elon Musk gave the only answer this week with an actual mechanism in it. CNBC reported him calling for cross-lab safety testing — labs reviewing each other's models before release — and he named Chinese labs as part of that. Whatever you think of the messenger, it's a procedure with named participants, which is more than the other two offered.

6. Damra Vol 00:02:58

   It's either the most serious proposal of the week or unworkable, and which one depends on what gets handed over. Weights? Evaluation results? A report someone wrote about a report? And it collides with the news cycle it's sitting inside. Anthropic spent these same days accusing Chinese labs of distilling its models through synthetic accounts. You can't argue for mutual model access and argue that access is being abused at scale, unless you're very specific about which access you mean — and nobody was.

7. Lenar Kess 00:03:29

There's a counterpoint from outside the chief-executive tier that's sharper than any of them. David Krueger wrote in the Guardian that Amodei's plan doesn't go far enough, and he put it plainly. Quote: what we need is a plan in which we're confident AI isn't going to kill us, and this isn't it. He anchors that to Jacob Coxon's resignation from Anthropic. And separately, Joscha Bach read the whole exchange as regulatory capture — his point being that you can have sincere safety staff and leadership that benefits from the rules those staff ask for, both at once, without anyone lying.

8. Damra Vol 00:04:06

   And then the Treasury Secretary walked into the same argument from a different door. Scott Bessent told House Financial Services he's against liability exemptions for AI companies. Hold the creators liable. That isn't deregulation and it isn't a pause. It's tort law, which is the oldest mechanism in the room and the only one that doesn't require anybody to agree on a definition of the word frontier before it can operate.

9. Lenar Kess 00:04:30

   So count the mechanisms. Huang wants self-regulation by whoever makes the product, Amodei wants a coordinated pause, Musk wants cross-lab peer review, and Bessent wants liability in court. Four incompatible answers inside one week, and the New York Times has been keeping a tally of two years of federal bills that went nowhere at all.

10. Damra Vol 00:04:50

    And Amodei told Al Jazeera he didn't appreciate how quickly AI would come to underpin the global economy. [pause] The person asking everyone to slow down says he misjudged the speed.

11. Lenar Kess 00:05:03

    Anthropic published a 154-page report on how people are misusing its models. Sourcing caveat first: this reaches me through a video walkthrough by Fireship rather than my own read of the document, and every number in it is Anthropic's own reporting on Anthropic's own platform. There is no outside auditor here. With that said, it sorts into seven categories. Four are the ones you'd guess — cyber warfare, influence operations, surveillance, and scams. The other three are biological misuse, conventional weapons development, and model distillation.

12. Damra Vol 00:05:37

    The cyber material carries the most specific detail. Midnight Blizzard, the Russian group, using Claude agents to rewrite malware that had already been flagged, so it stops matching signatures. A Chinese operation running what amounts to an autonomous zero-day foundry. And the one that stuck with me: 1.8 million Android application packages decompiled, looking for hardcoded application programming interface keys left behind in the binaries. It's a mechanical task at a scale

no human team would ever attempt, which is the kind of work these models are good at and the kind nobody wrote a threat model for.

13. Lenar Kess 00:06:15

    Then the distillation section, which names companies. Alibaba, DeepSeek, and Moonshot. Hundreds of millions of queries attributed across the three of them. Alibaba alone is put at more than a million requests a day, through synthetic accounts, to train Qwen. CONSTRUCT covered the accusation back in June. This is the documented version of it, with the volumes attached.

14. Damra Vol 00:06:37

    And the mechanics are worse than the accusation sounded, because distillation looks exactly like ordinary load on the API. Query in, answer out, and the bill paid on time. There's no packet you can block and no request that's wrong on its own. You catch it by account topology — signup patterns, payment sources, and query distributions that are too evenly spread across a curriculum. It's a fraud-detection problem wearing a safety problem's clothes, and the teams who are good at one are usually not the teams who own the other.

15. Lenar Kess 00:07:09

    Say more about why that bothers you.

16. Damra Vol 00:07:11

    Because of what Anthropic says about tiers. The report says distillation succeeded against Haiku, Sonnet, and Opus, but not against Fable or Mythos. No successful distillation observed is a statement about detection rather than a statement about what happened. And there's no method published. Nothing describes how they tell a distillation account apart from a heavy legitimate one. Whether the top tier resists distillation architecturally, or whether it's just harder to get access to in the first place, are two different claims, and that one sentence covers both.

17. Lenar Kess 00:07:45

    That points straight at export controls, and Matt Sheehan at Carnegie made the point in a New York Times question-and-answer this week. The controls reach hardware. They don't reach question-and-answer pairs coming out of a paid endpoint. Cloudflare published something adjacent alongside it — a proposal to split crawler identity, so a bot fetching a page to answer a live search query is distinguishable from a bot fetching that page to train on it.

18. Damra Vol 00:08:10

That's a serious piece of engineering, and it's pointed at the wrong leak. If the report is accurate, the models aren't being copied through crawlers at all. They're being copied through the metered endpoint, by accounts that look like customers and pay full price every month, on time.

19. Lenar Kess 00:08:26

    Microsoft has a safety paper that DAIR.AI summarized this week, and the attack in it is simple enough to describe in a sentence. You take an unaligned weaker model, have it split a harmful task into subquestions that each look innocuous, ask an aligned frontier model each subquestion in a separate session, and recombine the answers yourself. I only know this through the summary, so I'm not going to quote numbers at you that I haven't seen.

20. Damra Vol 00:08:51

    You don't need the numbers for the structural point. Every individual refusal works, and every single request is one the model should answer. The defense has to be cross-session state, and almost nobody keeps that — because for most providers the session is the unit of everything: rate limits, context, logging, and the refusal boundary itself. Asking a provider to reason across sessions is asking them to build a surveillance layer they've spent years promising customers they don't have.

21. Lenar Kess 00:09:21

    There's a second safety paper this week with actual numbers in it. It's called TAME — token attribution and masking for emergent misalignment. The finding is that emergent misalignment concentrates. The top five percent of tokens by attribution carry thirty-two percent of the attribution mass. Mask that five percent during fine-tuning and emergent misalignment drops by a factor of twenty-three in Llama and thirty-six in Qwen. Mask an equal number of random tokens instead and nothing changes.

22. Damra Vol 00:09:52

    What changed how I think about this is what those high-attribution tokens turned out to be. They were depleted for medical vocabulary and enriched for a register of unwarranted certainty. That holds controlling for rarity, so it isn't just rare words doing it. The fine-tune wasn't teaching the model bad medical facts. It was teaching the model a voice. A confident, unhedged way of asserting things. <emphasis>And the voice generalized to domains the training data never touched. </emphasis>

23. Lenar Kess 00:10:20

    Does the masking cost you anything?

24. Damra Vol 00:10:22

Some perplexity, and it stays concentrated on that same register, so you're not paying for it across the model's general fluency. That's what makes it usable rather than a curiosity. It's a preprint, run on released misalignment organisms, so anyone who wants to check it can check it directly instead of taking the authors' word.

25. Lenar Kess 00:10:43

    Yesterday I said we'd go read the six new filings in the consolidated OpenAI copyright litigation and tell you whether they contained exhibits or a dispositive motion. Answer: neither. There are three captions, all in the Northern District of California. M.G. versus Altman, A.C. versus Altman, and G.M. versus Altman. Each one posted the same document, and the document is a schedule. It sets deadlines for the defendants' motions to dismiss and a date for the initial case management conference. The defendants named are OpenAI Foundation, OpenAI OpCo, and OpenAI Group PBC, along with Samuel Altman personally.

26. Damra Vol 00:11:23

    Three captions, one lead case, identical text. That's related cases being put onto one calendar, not three new lawsuits — which matters, because a docket showing six filings in a day looks like escalation and is clerical. And I won't characterize what any of them claims from a caption. A caption tells you who is suing whom. It tells you nothing about what for.

27. Lenar Kess 00:11:47

    Separately, Sri Muppidi at the New York Times reports investors approaching OpenAI at a 1.2 trillion dollar valuation, with OpenAI's own position being at least 1.5 trillion. That's a report about talks, not a closed round. The number to hold onto is the gap. Three hundred billion dollars of disagreement about what the company is worth, unresolved and in public.

28. Damra Vol 00:12:10

    So OpenAI is arguing its floor is 1.5 trillion while its people sit in Canberra, telling Australian ministers that Australian copyright law is what's stopping them from training there.

29. Lenar Kess 00:12:22

    The Albanese government is considering a text-and-data-mining exception to Australian copyright law that would let AI companies train on copyrighted material by default, with an opt-out for rights holders. Josh Ittimani and Caitlin Cassidy at the Guardian have the leaked documents. It surfaced the same week senior OpenAI staff were meeting Labor ministers, and what sits on the other side of the trade is datacentre investment.

30. Damra Vol 00:12:48

Flipping a default changes who pays for doing nothing. Under the current rule, a photographer in Perth who does nothing is protected. Under the proposed rule, a photographer in Perth who does nothing is training data. Opt-out sounds symmetrical and isn't. It requires every rights holder in the country to find out the rule changed, then find the mechanism, then use it — on every platform, forever, and for free. Holly Rankin made that argument in a column the same day, at household scale, and it reads better than any of the industry submissions.

31. Lenar Kess 00:13:21

    Cloudflare's crawler split is the technical piece a rule like that would have to rest on. You can't write opt-out into law if nobody can tell which crawl is which. Their post was doing real numbers on Hacker News — sixty-seven points and forty comments — which for an infrastructure proposal means the people who'd have to implement it are already reading it.

32. Lenar Kess 00:13:41

    And the investment side of the trade isn't hypothetical. Anthropic signed a 31.9 billion dollar datacentre lease in western Queensland this week. Premier Crisafulli called it a major win for jobs and for the grid. Different company, different deal, no connection to the copyright proposal — but it tells you the going rate for what Australia is being asked to trade.

33. Damra Vol 00:14:03

    It's a leaked proposal, not a law. Senator David Pocock said easing copyright protections for datacentre investment would throw creatives under the bus, which is the sentence that gets quoted for the next six months. If it passes, Australia becomes the first place where the exchange — your training corpus for our construction jobs — is written down somewhere an ordinary person can go read it.

34. Lenar Kess 00:14:26

    Google shipped Gemini 3.8 Live, and a variant called Gemini 3.8 Live Extended Thinking. Mike Wheatley at SiliconANGLE covered it — real-time reasoning inside a speech model — which is trade coverage of a vendor announcement, so the claims are claims until somebody runs them. What I found more interesting is a talk that explains where the latency in these systems sits, and it turns out most of it isn't in the model at all.

35. Damra Vol 00:14:52

    Bohan Li gave that talk at AI Engineer, and there are three tricks in it. The first is a speculative transcriber. You run a fast streaming transcription model — Flux — alongside a slower, more accurate batch model called Scribe V2. The fast one gives you something to act on immediately, and the accurate one revises it as context resolves. Names and dates get corrected after the fact, which is

where a streaming model is weakest and where being wrong for four hundred milliseconds costs you nothing.

36. Lenar Kess 00:15:23

    The second one made me stop the video and go back. A background agent does tool calls while you are still talking. It does phonetic matching on names as you say them, fires the lookup before you've finished the sentence, and then injects the result into the main model's context.

37. Damra Vol 00:15:39

    And the main model behaves as though it made that call itself. [chuckle] There's a second half to it, too. The system runs eager generations, holds them in suspense, and cancels and retries them once the full utterance resolves. You're speculatively executing a conversation — branch prediction for dialogue. The third trick is the plainest of the three: a text-to-speech prefix cache. You keep about three words of pre-rendered audio for common openings, play those while Cartesia catches up over a websocket, and then suppress the overlap when the real audio arrives.

38. Lenar Kess 00:16:13

    Three words of cached audio is barely a trick. And it moves perceived latency more than most model upgrades would, which says something uncomfortable about what we've been measuring.

39. Damra Vol 00:16:23

    There's a paper doing the same move at the architecture level. It's called Never Stop Thinking, from Bojie Li and Noah Shi — continuous-time cognition out of an unmodified text model, by wrapping it in an orchestrator that interrupts and resumes. Nineteen percent off overall latency, and about half of that concentrated inside the regime they were targeting.

40. Lenar Kess 00:16:46

    A third paper in this neighborhood found that the standard way we evaluate these systems is backwards. They built ReactiveBench — a hundred and twenty scenarios, with pre-registered binary requirements and a verifiable streaming track, so the scoring doesn't depend on anybody's taste.

41. Damra Vol 00:17:03

    The judge reversal, yeah. Large language model judges reward visible reasoning. They see the model thinking and they score it higher. Put an independent judge on the same outputs and the advantage flips sign. And models trained against those judges complete fewer of the requirements when they think. The thinking was scoring well and performing worse, and the whole time the metric said it was working.

42. Lenar Kess 00:17:26

    So what's the fix?

43. Damra Vol 00:17:27

    Verifiable objectives instead of judged ones. Streaming completion goes from forty-eight percent to seventy-three, plus or minus five across seeds. That's under on-policy reinforcement learning, with a reward shaped by requirement type. It replicates at larger scale and on a second model family. And preference optimization couldn't get there — it could only trade one sub-goal against another. Two of the three wins in this segment come from the harness rather than the model. The speculative transcriber and the prefix cache would work the same way against a model from any vendor.

44. Lenar Kess 00:18:03

    There's a paper called Decomposition Buys Integrity, Not Yield, from Rong He, that fits a tree model against six hundred production deep-research traces. The headline number: each delegation tier retains about half of what the level below it found. Point five three six per tier, once off-target briefs are counted in. And one brief in sixteen goes off-target, measured across 1,012 annotated traces. Flat is optimal for yield. Every layer you add costs you half of what's underneath it.

45. Damra Vol 00:18:34

    What depth buys you instead is root-context exposure — N drops to N to the one-over-k. That's the trade, and it matters if your problem is that the root agent drowns in its own context. But there's a behavioral finding underneath that one too. Ask the model which sessions are worth delegating and it says somewhere between point seven and eleven point three percent, depending how you ask. The actual delegation rate in production is seven point eight percent. Close enough. Then they fit a hazard model over 743,819 production tool calls and found delegation doesn't respond to a filling context at all. It's an opening move. The agent decides to delegate at the start, before it knows whether it needs to.

46. Lenar Kess 00:19:21

    DAIR.AI had a complementary result from the same batch. Giving a model extra tools lowers its answer rate on questions that need no tool whatsoever. The tools aren't being used badly. Their presence alone changes the behavior.

47. Damra Vol 00:19:36

    And then Sara Vera Marjanović and colleagues have one called Mo' Models, Mo' Problems, which tests eight model-selection strategies and finds that expanding your pool often drops you below the

best single base model you started with. Same-family selection wins. Heterogeneity was the source of the instability, not the diversity benefit everyone assumes it is.

48. Lenar Kess 00:19:58

    Three papers in one week pointing the same direction: fewer agents, fewer tools, fewer model families. [pause] That's not what anybody's product roadmap says.

49. Damra Vol 00:20:09

    On the runtime side, there's a single-author preprint called JustFit with self-reported numbers I'd want somebody else to reproduce before I leaned on them. It's an Apple MLX runtime running a 27-billion-parameter Qwen model on a 24-gigabyte M4 Pro MacBook. It fits 196,608 input tokens and 16,384 of output. The mlx-vlm baseline manages 30,720 positions on the same hardware, so call it roughly seven times. About nineteen tokens a second on a 32-thousand-token probe, and it solved 29 of 30 AIME 2026 problems.

50. Lenar Kess 00:20:53

    And there's a compiler paper called Ave that checks data-flow invariants at compile time. The compiler propagates symbolic tags through the program, and an SMT solver — a satisfiability checker — verifies the assertions at each use site. They ran it on AMD MI300X across matrix multiply, flash attention, and mixture-of-experts layers. Those three together account for up to ninety percent of GPU time in inference. They got eighty-nine to ninety-nine percent of hand-optimized library throughput.

51. Damra Vol 00:21:26

    What I'd use it for is the counterexample. Violate an invariant and the solver hands you back an actual input that breaks it, instead of a test that failed for reasons you then have to go find.

52. Lenar Kess 00:21:37

    What else did you have?

53. Damra Vol 00:21:39

    Two quick ones. TypeSafe AI raised forty million dollars for something called Jev: typed probabilistic decisions, which Diogo Almeida describes as reinforcement learning for calibrated decisions. Thomas Claburn wrote it up at The Register, and there are no independent calibration numbers anywhere in it, so treat it as a company claim about a company product. SemiAnalysis also has a piece on datacentre permitting with the number I've been wanting. There are three hundred-plus local moratoriums and a New York executive order. Twenty gigawatts of capacity sits inside

some restricted boundary, and the amount that slips is 1,525 megawatts. Call it 2.3 gigawatts nationwide.

54. Lenar Kess 00:22:23

    That puts the pacing argument on different ground. Local opposition isn't what's constraining capacity, even though NBC was in Mason, Michigan this week watching a datacentre fight turn into a midterm issue. Anthropic leased 31.9 billion dollars of new capacity in western Queensland, and hired a Head of Frontier Compute Strategy to go build coalitions for more. Nothing physical is slowing this down. Which means every brake left in the story is one somebody would have to choose to pull — and the only person this week who named a mechanism with a date attached to it was a federal judge setting a briefing schedule. That's Braid for Wednesday, September sixteenth. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic