The boundary is an instruction

2026-08-21 / 00:23:52

> *“If you decided right now that an agent should stop having access to your billing system, how many seconds until that's true?”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Anthropic took computer use, Skills, and Files to general availability on the same day two conference talks argued that an agent's authority has to be enforced somewhere the model can't reach. Meanwhile a published Claude artifact showed up in Google results as a fake install page, and the revenue number behind a rumored IPO got picked apart in public.

- Claude Developers on the GA release — computer use, the browser tool, the Skills API, and the Files API all generally available, plus a toolset dated the first of August that batches click, type, key, and screenshot into one turn. Anthropic's own early-access figure is twenty to forty percent fewer round trips.
- A single unconfirmed r/ClaudeAI post describing a macOS infostealer reached through a search result that was a published artifact on Anthropic's own domain. One post, zero votes, no forensic writeup — but the structural part is checkable.
- Sarthak Aggarwal's sixteen-minute talk on planner-executor separation, scoped tokens bound to actor and expiry, and the number nobody publishes: revocation latency.

- Docker's portable-runtime talk — micro-VMs, stubbed credentials, network policy from outside, capabilities composed just-in-time over MCP.
- Payman Connect gives an agent bank access, and Yohei Nakajima asked the obvious question the same afternoon.
- The AI Daily Brief on the Bloomberg IPO report, Dylan Patel's methodology objection to the sixty-five billion ARR figure, and the super-voting share structure.
- DeepSeek-V4-Flash-Vision-Exp went API-only, alongside a harness where the control loop itself is a plugin. Fireship's single run is the only cost anchor: twenty-nine minutes, 2.6 million output tokens, thirty dollars.
- Dan Bjornn on the calcification tax — why Lease End tore out a fine-tuned model that made twelve million dollars in a year.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Say you're setting up a new Mac this morning and you want Claude Code on it. You do what everyone does. You type install Claude Code into Google, you take one of the first results, it looks

like the docs you've seen before, and you follow the instructions. [pause] Before you paste that command into a terminal, what's the check you actually run? If we're being accurate about our own behavior, the check is the domain. You glance at the address bar, you see a name you recognize, and that's the end of the audit. A user on the ClaudeAI subreddit posted yesterday saying they ran exactly that sequence. What came out the other end was a macOS infostealer.

- x.com
- x.com
- x.com
- x.com
- reddit.com
- youtube.com
- youtube.com
- x.com
- x.com
- v.redd.it
- x.com
- youtube.com
- x.com
- x.com
- youtube.com
- openrouter.ai
- x.com
- x.com
- youtube.com
- youtube.com
- openai.com
- x.com
- x.com
- x.com
- x.com
- mathstodon.xyz
- blog.curiousquail.com

2. Damra Vol 00:00:41

   And here's what makes it more than a normal phishing story. The page was a published Claude artifact, sitting on Anthropic's own domain. So the check people were taught to run returned a green light, correctly, because the domain really was Anthropic's. User-generated content inherited a vendor's reputation, and Google's ranking did the rest of the work.

3. Lenar Kess 00:01:03

   Let's be precise about what we know, because this is one post with zero votes and no independent confirmation. The poster describes the chain. Search, first page, a page that looked like install docs, a command, and then a machine that started leaking credentials. We're not going to name or reconstruct the command, and nobody has published a forensic writeup. What's checkable is the structural part. Anthropic does host user-published artifacts on its own domain, and a search engine will happily rank one of them.

4. Damra Vol 00:01:34

   Right, and that's a design decision anybody who runs a publish-your-thing feature has made at some point. You put user content on your apex domain because it's simpler, because subdomain isolation is a pain, and because your CDN config already exists. Then one day the signal your users trust most about you turns out to be one you handed to a stranger. [tsk] As an attack it's old. What's newer is the population being targeted — people installing agent tooling are exactly the people who will run a shell command without reading it closely.

5. Lenar Kess 00:02:07

   That story sets up most of the rest of the episode. Anthropic took computer use, Skills, and Files to general availability yesterday, along with a new toolset that batches multiple actions into a single turn. Then two conference talks and one payments launch all circle the same question — what is an agent allowed to do, and who can revoke it. Bloomberg says Anthropic could file to go public by the end of this month. Meanwhile the revenue number that would anchor that offering is being disputed in public. DeepSeek shipped a vision model and an open coding harness on the same push. And a data scientist got on stage to explain why he ripped out a fine-tuned model that had made twelve million dollars.

6. Damra Vol 00:02:49

   Which is a better lineup than it sounds. Two of those are the same story told from opposite ends. One side is shipping more agent authority. The other side is standing at a conference explaining that authority without a boundary is a hope, not a design.

7. Lenar Kess 00:03:04

   Start with the artifact. The Claude developer account posted yesterday that computer use, the browser tool, the Skills API, and the Files API are all now generally available on the Claude platform. There's also a new toolset version, dated the first of August, that lets the model issue several actions in one turn. It can click and type and hit a key and grab a screenshot as a single batch, rather than paying a round trip for each one.

8. Damra Vol 00:03:31

   That's the detail that changes the arithmetic. The old loop was brutal. Take a screenshot, send the whole image up, wait for the model to decide to click at a coordinate, send the click, take another screenshot. Every single interaction paid a full network hop and a full image of context. If you're driving an application that has no API — which is most enterprise software written before 2015 — you were paying model latency for something a macro used to do for free.

9. Lenar Kess 00:03:59

   Anthropic says early-access customers saw twenty to forty percent fewer round trips per task. That's their own number from their own early-access program, so take it as a vendor claim rather than a measurement. Directionally it's the obvious win. I'd expect the real effect to be larger on cost than on wall-clock time, because the screenshots are the expensive part of the context.

10. Damra Vol 00:04:22

    Here's what nags at me though. Every batched action is one fewer place you can put a check. If the model emits click, type and key as one unit, then whatever sits between the agent and the machine sees a plan rather than a sequence of individual decisions. You can still gate the batch. You just can't gate action three based on what the screen looked like after action two — which is the exact moment where a page you didn't expect shows up.

11. Lenar Kess 00:04:48

    That's a real trade and I don't think it's a fatal one. It's what you buy with any batching layer: throughput, in exchange for a shorter feedback loop. What I don't know yet is whether the policy layer people build around this is expressive enough to reason about a whole batch. Which is why the conference talks from yesterday are interesting rather than academic.

12. Damra Vol 00:05:08

    Before we leave the platform news — Boris Cherny posted about Mythos-class deployments where the enterprise customer owns and controls their own data under their own compliance rules. That sounds like boilerplate until you put it next to the fact that Anthropic is reportedly weeks from an S-1. Enterprise data-residency language is what you write when procurement teams with actual lawyers are the ones deciding.

13. Lenar Kess 00:05:33

    And Managed Agents got paired with the AG-UI protocol from CopilotKit in the same window, which we mentioned yesterday. I'll leave it as a mention again. What catches my attention is that the interface layer for agents is converging on someone else's protocol rather than a proprietary one.

14. Lenar Kess 00:05:50

    At the AI Engineer conference yesterday, Sarthak Aggarwal, co-founder of Deca Work, gave a sixteen-minute talk whose central claim is very simple. A code freeze that exists only as an instruction isn't a boundary. He uses two incidents to make the case. One is the Replit production-deletion event, where an agent went past a stated freeze and deleted production data. The other is the EchoLeak vulnerability in Microsoft 365 Copilot — a zero-click chain where an external email became context and forced data exfiltration downstream, with nobody clicking anything.

15. Damra Vol 00:06:27

    His actual proposal is more interesting than the incidents. He wants agents to carry something like an operational identity card. Who the actor is, who they're acting on behalf of, and what the delegation context is. Then the exact capabilities granted, the policy that governs them, and how long revocation takes. That last one is the item nobody publishes. Revocation latency. If you decided right now that an agent should stop having access to your billing system, how many seconds until that's true?

16. Lenar Kess 00:06:59

    His answer to what's missing structurally is that OAuth gives you part of it and stops short. There's no standardized on-behalf-of identity model for a software actor that was itself dispatched by another software actor. The architecture he arrives at is planner-executor separation with a hard policy gate in between. The trusted intent gets normalized into a typed, logged plan before the system ever touches untrusted evidence. Then the executor runs approved actions against short-lived scoped tokens bound to actor, subject, audience, and expiry. And it never gets to re-read the original context.

17. Damra Vol 00:07:37

    That last constraint is the clever bit and it's easy to skim past. The executor can't look back at the email, the web page, or the document. It only sees the approved plan. So an injected instruction living in the untrusted text has no surface left to talk to. You're not asking the model to resist temptation. You're removing its ability to hear the request.

18. Lenar Kess 00:08:00

    Docker's talk the same afternoon comes at the same gap from the containment side. The argument there is that intelligence stopped being the bottleneck and safe autonomy became it, because agents expand their own goals at runtime and a static permission set can't anticipate what they'll need. What they propose is a portable runtime. Micro-VMs, with the security controls enforced outside the agent boundary. Credentials injected as stubs rather than real secrets. Network policy set from

outside. And capabilities composed just-in-time over MCP, so the agent gets an exact data subset instead of read access to a whole store.

19. Damra Vol 00:08:39

    They demoed a tool called SPX doing it, and the portability claim is the ambitious part. Same constrained sandbox on your laptop, in their cloud, or in a customer's private cloud, migrated with a flag. That's the Docker pitch restated as a safety pitch. Fine — they've earned that analogy. Whether the intent-evaluation layer works is a separate question, because approving or escalating a capability request based on whether it matches user intent is itself a judgment call made by a model.

20. Lenar Kess 00:09:10

    So that's the mechanism side. Now put the money side next to it, on the same day. A company called Payman shipped Payman Connect, which gives an agent access to a person's bank account. Yohei Nakajima's reaction was essentially to ask the obvious question in public, which I appreciated more than a hot take.

21. Damra Vol 00:09:28

    [chuckle] And underneath that, on the ClaudeAI subreddit, somebody posted a month of letting Claude manage a real chunk of money. The result was a thirty-one thousand dollar loss. Self-reported, no audited statement, so treat the number as a claim. But the two arriving within hours of each other is the state of play. The talks describe a delegation architecture nobody has standardized, and the products are already handing over the bank.

22. Lenar Kess 00:09:56

    The gap I keep coming back to is auditability as an operational tool rather than a compliance artifact. Aggarwal's point is that the audit trail isn't there so you can pass a review. It's there so that when something goes wrong at two in the morning you can answer which agent, acting for whom, under which grant, took the action — and then kill that grant specifically instead of turning off the whole system.

23. Damra Vol 00:10:20

    That's a different ask from how logging usually gets built, which is: write everything, search it later, and hope the field you need is in there. He's asking for logs that are structured enough to revoke from. Nobody ships that on version one.

24. Lenar Kess 00:10:34

Bloomberg reported yesterday, relayed through the aggregator accounts, that Anthropic could file for an IPO as soon as the end of August, with expectations it would match or exceed SpaceX's record. That's the headline. The more useful development on the same day is that the revenue number which would anchor that offering got attacked on methodology.

25. Damra Vol 00:10:56

    Dylan Patel at Semi Analysis. His objection is specific enough to check, so let me repeat it exactly. Anthropic's claimed sixty-five billion in annual recurring revenue is, per Patel, an extrapolation of four weeks of API revenue. Take a month, multiply by thirteen, call it recurring. API spend from a coding-agent boom is about the least recurring revenue that exists. It moves with whatever model people are excited about that month.

26. Lenar Kess 00:11:24

    His second objection is the one I hadn't thought about carefully. Semi Analysis says over forty percent of recent annual recurring revenue comes through indirect hyperscaler channels like Bedrock and Foundry, where the revenue is being counted gross. So the platform's cut sits inside the number that's being presented as Anthropic's revenue.

27. Damra Vol 00:11:44

    Which is a normal accounting fight rather than a scandal, and it'll be settled by an S-1 rather than by anybody on the timeline. But it does mean the sixty-five billion figure that's been repeated everywhere for weeks is going to look different once it's in a filing with an auditor's name attached. That's usually where these numbers get smaller.

28. Lenar Kess 00:12:03

    The governance piece is the other item. Anthropic plans to issue super-voting shares giving Dario Amodei and the co-founders board control on roughly fifteen percent of the equity. Structurally that's Google, Meta and Snap. It's the standard founder-control template for a tech listing.

29. Damra Vol 00:12:21

    It is standard. I think the reason it draws more attention here than it did for a social network is that Anthropic's whole public argument is that this technology is societally consequential and needs careful stewardship. If your pitch is that judgment matters enormously, people are going to look harder at who holds the votes. That's simply the cost of making the argument.

30. Lenar Kess 00:12:43

For contrast on the other side of the street. OpenAI reported a forty billion annualized run rate and six point seven billion in second-quarter revenue, growing eighteen percent, with operating margins still negative. They also put a fifty percent token discount on GPT-5.6 through OpenRouter and the Vercel gateway. Semi Analysis reads that as a targeted marketing move against Chinese models and Anthropic rather than a general price war, and it did what it was supposed to. The Luna variant became the top closed model on OpenRouter.

31. Damra Vol 00:13:16

    Discounting on the one platform where the leaderboard is public and the press checks it daily is a very cheap way to buy a narrative. People are running those tokens, sure. But the metric being moved is a niche one, and it gets reported as though it were market share.

32. Lenar Kess 00:13:30

    Elsewhere. DeepSeek put DeepSeek-V4-Flash-Vision-Exp live on their API platform this morning. It's an experimental multimodal model they say holds V4-Flash's text performance, including agents, reasoning, and world knowledge. API-only right now. Prince Canuma picked it up within twenty minutes, which is usually the leading indicator that somebody's going to try to run it locally.

33. Damra Vol 00:13:54

    Open weights haven't been announced, though, and I'd rather say that plainly than let the anticipation do the work. The other half of the push is the more architecturally interesting one. The DeepSeek Harness is a coding agent framework where basically everything is a plugin. That covers the model adapters and the tool integrations, the sandboxes and the user interface, and — this is the unusual part — the core control loop itself. You can swap any of them, and you configure it in YAML.

34. Lenar Kess 00:14:24

    Making the control loop a plugin is a real statement. That's the piece every closed coding agent treats as the product. If you can hot-swap the loop, the harness has stopped claiming to know the right way to run an agent. It's saying the right way will keep changing, and you shouldn't have to fork anything to track it.

35. Damra Vol 00:14:41

    There's also a trajectory panel that logs reasoning and tool calls the way a stack trace logs frames. Which, if it works, is the debugging surface people have been hand-rolling for a year. And the whole thing is model-agnostic — you can point it at an external inference provider. DeepSeek shipped a harness that helps you not use DeepSeek.

36. Lenar Kess 00:15:02

    Fireship ran it and gave us the only cost anchor anybody has. DeepSeek V4 Pro at maximum settings, in standard mode, worked for twenty-nine minutes and burned two point six million output tokens for thirty dollars. What came out was a working Node and React app with swipe animations and chat. His read on quality was competent, though less polished on the interface than what he gets from Codex or Fable.

37. Damra Vol 00:15:28

    That's one test, one app, and one person's afternoon — an anecdote with a receipt attached rather than a benchmark. But thirty dollars for twenty-nine minutes of autonomous work is the number I'd want in my head, because it tells you the unit you're now budgeting in. Not tokens per query. Dollars per attempt.

38. Lenar Kess 00:15:47

    Two quick adjacent items. A stealth model called Ox Alpha showed up on OpenRouter's stealth endpoint and hit the Hacker News front page at a hundred sixty-eight points, with a community claim that it beats Fable on software-engineering tasks. Unverified. And the only fingerprint anybody has for who built it is which questions it declines. Commenters report it refuses on Tiananmen Square while answering electronic-warfare questions that Opus and Fable turn down.

39. Damra Vol 00:16:14

    Which is a strange way to do attribution, and it works better than it should. The refusal profile is the one part of a model you can't easily hide behind a stealth endpoint, because it's baked into behavior rather than metadata. I wouldn't state an origin. I'd just note that the safety training has become a fingerprint.

40. Lenar Kess 00:16:33

    And Grok 4.6 scored fifty-nine on the Artificial Analysis Agentic Index, tied for first with Claude Opus 5 at max settings. Musk posted it. The more durable detail from the same afternoon is that the Midas Project's watchtower account noticed xAI revised the Grok 4.6 model card after publication, with an incomplete changelog.

41. Damra Vol 00:16:57

    A model card that changes after people have already read it, without a full record of what changed, undermines the one artifact that's supposed to be stable. [tsk] The benchmark number will be reproduced by somebody within a week. The version of the card you read last Tuesday is gone.

42. Lenar Kess 00:17:14

    This one's my favorite artifact of the day. Dan Bjornn, a senior data scientist at Lease End, gave a talk about a customer-messaging system they deployed in late 2024 for lease buyout inquiries. It started as workflow-driven retrieval-augmented generation, with a vector database and intent classification across six categories. Then they moved to supervised fine-tuning, chasing accuracy, lower latency, and cheaper inference on a smaller model. It produced twelve million dollars in revenue at fifty times return in a year. And they tore it out.

43. Damra Vol 00:17:49

    The failure modes he describes are the good part, because they're specific and slightly absurd. A customer replies good morning, and the system triggers an immediate phone call. Someone confirms an appointment for next week and the model reads it as a request to be contacted right now. Call it a classifier that learned a pattern from the data and then applies it with total confidence.

44. Lenar Kess 00:18:12

    And the fix cycle was about a week each time. Collect data, label it with a model as judge, validate by hand, then deploy. Then the next regression shows up somewhere else. He calls the accumulated cost a calcification tax. The training data you produced is provider-specific, so you're locked in, and the architecture is frozen at the patterns that made sense in late 2024.

45. Damra Vol 00:18:36

    I'd underline that for anyone who fine-tuned something eighteen months ago and has been feeling vaguely good about it. The lock-in doesn't live in the weights. Your labeled dataset was shaped for one provider's training format, and the workflow around it was designed for a paradigm that got replaced while you were maintaining it.

46. Lenar Kess 00:18:54

    The rebuild replaced weights with dynamic skills, tools, and context injection. Problem-to-deployment went from roughly a week to under an hour. Per-message API cost went up, because they're calling bigger frontier models now. Total cost went down, because the engineering overhead disappeared. Accuracy improved substantially. Latency, he says, barely moved — which is the admission that makes me trust the rest of it.

47. Damra Vol 00:19:20

    And he keeps a carve-out that deserves repeating. Fine-tuning still wins when you need offline deployment or strict data privacy. His claim isn't that the technique is dead. It's that for production intent classification, paying more per call to stop maintaining a training pipeline is the better trade.

48. Lenar Kess 00:19:38

    The companion talk in the same session is Niels Rogge from Hugging Face's community science team, who automated the job of nudging researchers to publish models and datasets on the Hub instead of Google Drive or Zenodo. A nightly cron job through GitHub Actions scans arXiv papers and their repos, then opens issues and pull requests. Hundreds of issues a night. Two negative responses out of thousands.

49. Damra Vol 00:20:03

    And researchers from Apple, DeepMind, and the Paddle OCR team actually moved their artifacts because of it. But there's a choice in there he states openly, so we should say it plainly as well. He deliberately doesn't disclose that the outreach is automated, because disclosure would change how researchers engage. That's a defensible product decision, and it will also feel different when everyone does it.

50. Lenar Kess 00:20:27

    He also moved from a deterministic pipeline to actual agents for the follow-up phase, and switched from the Claude Agents SDK to GLM 5.2 through Hugging Face inference providers. Same task, model swapped underneath, because the abstraction let him. Which is the exact argument Bjornn is making from the other direction.

51. Lenar Kess 00:20:47

    Last stretch. OpenAI launched a blog called AI Futures yesterday. The first post is by Dean Ball, and its subject is what he calls the most challenging risk area in AI policy — concentration of power.

52. Damra Vol 00:21:01

    Then Nathan Calvin and Miles Brundage spent the same day pushing back on comments OpenAI made to Bloomberg about state AI auditing requirements. Calvin's specific objection is to the company's analogy. OpenAI compared Illinois audit requirements to checking the brake lights, and Calvin rejects that as a description of what the rules actually ask for.

53. Lenar Kess 00:21:23

    And Sneha Revanur, who had read OpenAI's endorsement of Illinois Senate Bill 315 as a sign that the company's policy engagement had settled into something constructive, says this fight looks like the older pattern. I'm going to leave both facts where the sources leave them. A company can publish serious work on concentrated power and also fight an audit rule it thinks is badly drafted. Those are two different conversations, and collapsing them into one costs you the ability to evaluate either.

54. Damra Vol 00:21:52

    What I'd want from the AI Futures blog is whether it engages with the audit question at all. Or whether concentration of power stays safely at the level of nation-states and frontier labs, while the concrete mechanism for checking a lab's behavior gets argued about somewhere else entirely.

55. Lenar Kess 00:22:08

    One more, briefly, because it changes a commercial question for a lot of people listening. Hacker News spent yesterday on a post at a hundred forty-two points and a hundred twenty-five comments, arguing that copyright doesn't protect AI-generated content in the EU. The source is a Mastodon post rather than a ruling or a statute, so I'd hold the confident version of this.

56. Damra Vol 00:22:30

    But the thread keeps circling something that isn't going away. How much human contribution restores protection, and who gets to decide that. If a meaningful share of your codebase or your creative output is model-produced, the ownership question stops being philosophical.

57. Lenar Kess 00:22:46

    There's a second thread next to it at fourteen hundred points, about Aaron Swartz being prosecuted for scraping while Meta does it in bulk without consequence. That one runs on anger rather than law, and the anger is earned. I'll note the asymmetry it names. The same act reads as a felony or as a data pipeline depending on who performs it.

58. Damra Vol 00:23:07

    Which brings the day around to what I'd check first tomorrow. Anthropic moved agent control to general availability and shipped batching that makes each turn do more. Two talks the same afternoon argued that authority has to be bounded outside the model, with revocation you can actually measure. Nobody has published a revocation latency number yet.

59. Lenar Kess 00:23:29

    That's the number I'd like to see in a platform document rather than a conference slide. Until then, the accurate description of most agent deployments is that the boundary is an instruction, and the instruction is a hope. Aggarwal's talk runs sixteen minutes, and out of everything from yesterday it's the one I'd spend the time on.

Hosts on this episode

- Lenar Kess moderator

- Damra Vol critic