Someone Else's Retry Logic

2026-08-09 / 00:25:49

> *"You cap a channel, and traffic finds the uncapped one next to it. The file name is the uncapped one."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Two vendors shipped managed agent runtimes on the same Saturday, Claude Code sessions can now message each other over a socket, and one developer found his review agent had burned six thousand turns supervising two hundred of real work. The tension across all three: the layer nobody wants to own is the layer that decides what everything costs.

- Harrison Chase's three-layer stack — model, harness, runtime — arrives the same day Anthropic publishes on managed agents, which means the seam in the middle is now something you buy rather than build.
- Anthropic's cross-session messaging docs say plain text only, which explains why agents have been base64-encoding attachments into file names: they're routing around a deliberate cap, not inventing a protocol.
- Faxan's usage audit found Codex auto-review at roughly 6,000 turns in a day against 224 coding turns, and nothing alerted him.
- 404 Media on SAP reports most travel and hiring frozen over AI spending, with AI-related travel and hiring exempt — a token bill competing with headcount.

- Miles Brundage names the narrow path between a self-regulatory body with no safety floor and rules applied selectively by whoever has access.
- A tokenizer comparison put the same 330 lines of code at 1,609 tokens for Qwen and 4,258 for Gemma, which is a cost multiplier hiding in a design choice nobody advertises.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Here's what I keep circling back to. Say you're building with agents right now, and I mean an actual thing that runs on a schedule and touches a real database. How much of what you've written is the interesting work? My guess is the ratio is bad. You've got retry logic, sandbox setup, credential handling, and some hand-rolled way of remembering what happened on the last run. That's most of the job. And yesterday, on a Saturday, two of the biggest names in agent tooling both shipped a product whose pitch is: stop writing that, rent it from us.

   - x.com
   - x.com
   - x.com
   - x.com

- x.com
- x.com
- x.com
- code.claude.com
- x.com
- x.com
- x.com
- reddit.com
- x.com
- marketnow.site
- x.com
- x.com
- reddit.com
- reddit.com
- 404media.co
- i.redd.it
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- reddit.com
- reddit.com
- reddit.com
- arstechnica.com

2. Damra Vol 00:00:37

   Two of them, on a Saturday, which already tells you something about who was watching whom. Anthropic published an engineering post on managed agents — the line people keep pulling out is separate the brain from the hands. And LangChain shipped managed deep agents, which is their harness with LangSmith infrastructure underneath it. Harrison Chase spent most of yesterday posting about the same idea from three different angles.

3. Lenar Kess 00:01:02

   Give me his version of the stack, because the convergence shows up there.

4. Damra Vol 00:01:06

He draws three layers. The model sits at the bottom. The harness sits in the middle, and that's the loop, the tools, and the planning. Then a runtime wraps around the whole thing, holding durable execution, sandboxes, authentication, and memory. He argues that a roughly standard version of that stack has emerged, and a managed platform is just someone packaging the layers you'd rather not own.

5. Lenar Kess 00:01:29

   And the pushback underneath his post amounts to: you've drawn the hard part as one box.

6. Damra Vol 00:01:35

   Caspar Broekhuizen said it plainly — productionization is where the difficulty lives and where the cost decisions get made. He means auth, sandboxes, and durable execution. Those aren't components you drop in. Each one encodes a set of defaults about what your agent may touch and how long it's permitted to keep trying.

7. Lenar Kess 00:01:55

   That second half is the money question. How long it keeps trying is a billing decision dressed as an engineering decision. If I rent the runtime, I've handed the vendor the dial that controls my invoice.

8. Damra Vol 00:02:07

   Right, and you've handed them the dial that decides when your agent gives up, which is also a correctness decision. There's a reply from Raja Meer Baz making a sharper version of that. His read is that the contested territory sits between the bottom layer and the top one. Whoever owns the seam in the middle sets the terms for everybody standing on either side.

9. Lenar Kess 00:02:27

   Does that hold up, though? It sounds like a nice abstraction, and I'd like to know whether it survives contact with anything.

10. Damra Vol 00:02:33

    Partly. It holds for auth and sandboxing, where there really is a boundary someone has to define and enforce. I'm less convinced about memory. Sena raised that in a reply — whether memory is a standard component of the stack at all, or whether it's application-specific enough that packaging it is a category error. Chase posted separately about memory in managed deep agents, so they're betting it generalizes.

11. Lenar Kess 00:02:58

My instinct is it doesn't generalize, and the reason is that memory is where your product's opinions live. What an agent chooses to remember about a user is the difference between two products built on identical models. Handing that to a platform means handing over what made you distinct.

12. Damra Vol 00:03:15

    Unless you're a team of four and you were never going to build a good memory layer anyway, in which case a mediocre managed one beats nothing. [chuckle] That's the split, though. The people for whom this is a bad trade are the people who could afford not to take it.

13. Lenar Kess 00:03:30

    Let me name the caution before we move on, because this is easy to overread. Two vendors shipping similar products within a day of each other is convergence, not a standard. Nobody has agreed on anything. There's no spec and no interop story, and if you build on LangChain's managed runtime you aren't one config change away from Anthropic's.

14. Damra Vol 00:03:52

    And there's a much older pattern underneath it. This is the same arc as hosted databases and hosted queues. First everyone rolls their own, then the design stabilizes enough that renting it becomes the right call, then five years later the interesting companies are the ones who took the layer back in-house for a specific reason. We're at step two, and step two is usually right.

15. Lenar Kess 00:04:14

    Okay. Here's where we're going today. That's the commercial story about agent infrastructure. There's a separate capability story — Claude Code sessions can now message each other, and I read the documentation this morning and found something that complicates the version circulating online. Then there's an operations story, which is Codex burning six thousand turns in a day reviewing its own work. After that we've got the reported reason Demis Hassabis stayed at Google, SAP freezing hiring over its AI bill, and Miles Brundage on what the law fails to do. Then a few smaller measurements that are better than they sound.

16. Damra Vol 00:04:52

    Three agent stories that aren't the same story, and I'll flag that now, because they happened the same weekend and it'll be tempting to weld them together.

17. Lenar Kess 00:05:01

    So Anthropic documented cross-session messaging for Claude Code. The premise is simple: one session on your machine can send a piece of text to another session on your machine. You run a

slash command — list-agents — to see who's reachable, and under the hood Claude has one tool for listing the sessions it can see and another for delivering the message. It requires version 2.1.224 or later, and it runs on macOS and Linux but not on native Windows.

18. Damra Vol 00:05:30

    And the delivery mechanism is lovely. Each session binds a Unix domain socket on disk, restricted to your operating system user. Same-machine messages never touch Anthropic's servers — they go over that socket. You can see the address yourself; it shows up in the status output as a peer address row.

19. Lenar Kess 00:05:48

    Which means the whole thing is filesystem-scoped. Two sessions can only find each other if they can see the same files.

20. Damra Vol 00:05:55

    Exactly, and the docs spell out the consequence: a session inside a container and a session on the host can't reach each other, because the container has its own filesystem. Two sessions inside the same container can. That's a nice property — the reachability boundary is something you already understand, rather than a new concept you have to learn.

21. Lenar Kess 00:06:16

    Now here's a correction, because the version going around is off. Simon Willison flagged the strange detail that some of these agents coordinate through file names alone — base64-encoding attachments into the name and prefixing it with z-z so the message sorts to the bottom of a directory listing. That got read as: look at the weird improvised protocol the agents invented.

22. Damra Vol 00:06:40

    [tsk] And when you read the actual documentation, the improvisation makes complete sense. Under limitations, it says plain text only. Claude sends only plain text across sessions. Structured protocol messages stay inside an agent team. So there's no attachment channel, because Anthropic deliberately didn't build one.

23. Lenar Kess 00:07:00

    So the agents aren't inventing a protocol out of nowhere. They're routing around a constraint the product imposed on purpose, using the one shared surface that's left.

24. Damra Vol 00:07:09

Which is the oldest move in computing. You cap a channel, and traffic finds the uncapped one next to it. The file name is the uncapped one. <emphasis>Nobody</emphasis> put a size limit on file names.

25. Lenar Kess 00:07:21

    [laugh] Nobody ever does.

26. Damra Vol 00:07:23

    And I'd say the permission design here is more thought-through than the reaction suggests. A message from another session cannot approve anything — it never counts as your consent, so it can't answer a pending permission prompt. It can't change permission settings or your project instructions file. And a slash command inside the message text arrives as plain text; it never executes.

27. Lenar Kess 00:07:46

    What decides whether a message gets through at all?

28. Damra Vol 00:07:48

    There's an inbound-message setting with three positions — accept, hold, or refuse. When nothing's configured, the default sorts sessions into two classes: the ones that prompt for permissions, and the ones that bypass prompts. If your receiving session bypasses prompts, every incoming message gets held for your approval unless the sender also bypasses. And a held message expires after five minutes by default and gets dropped.

29. Lenar Kess 00:08:15

    So the strictest configuration is the one that asks you the most, which is backwards from how people will expect it to feel.

30. Damra Vol 00:08:22

    It's backwards from how it feels and correct in how it behaves. A session with no prompts is the one you least want accepting instructions from elsewhere. There's also loop throttling written in — identical repeats inside a short window get dropped, and there's a cap of fifty accepted messages waiting to be read. The docs say a message loop between two sessions stops on its own.

31. Lenar Kess 00:08:45

    Someone thought about the runaway case before shipping, which given the next segment is going to look almost pointed.

32. Damra Vol 00:08:52

    Meanwhile Alon Wolenitz shipped pi-claude-link, an open-source extension that puts Pi sessions onto the same mesh. Which is the predictable next step — the moment there's a socket and a naming convention, somebody bridges another agent onto it.

33. Lenar Kess 00:09:08

    The top comment on the Hacker News thread makes the uncomfortable comparison directly, and I'll attribute it carefully, because it's a commenter's connection rather than a documented one. They point at the Hugging Face incident and note that the first thing that swarm did was build itself a messaging system. We spent a lot of time on that incident yesterday so I won't relitigate it, but the observation stands on its own: agents keep reaching for coordination, and now there's a supported way to do it.

34. Damra Vol 00:09:38

    With rate limits, a consent model, and a socket you can point at. I'd rather have the version with the approval dialog than the version they improvise at three in the morning, and there's a strand of the reaction that seems to want neither.

35. Lenar Kess 00:09:50

    Faxan went looking at a usage graph that didn't make sense and found something specific. Codex's auto-review system had consumed close to six thousand turns in a single day. The actual coding agents, the ones doing the work he asked for, accounted for two hundred and twenty-four. So roughly a twenty-five to one ratio between reviewing and doing, and his read is that it was stuck in an approval-state loop.

36. Damra Vol 00:10:14

    Twenty-five to one. And the arresting bit isn't the waste — it's that the loop stayed invisible until he audited it. Nothing alerted him, and nothing broke. The system sat in a state where it kept deciding the work needed one more look, and the only symptom was a graph that looked wrong.

37. Lenar Kess 00:10:31

    I'd be careful not to inflate one person's audit into a claim about how OpenAI bills people. This is a single account over a single day.

38. Damra Vol 00:10:40

    Agreed, and it's still the most concrete report I've seen of a supervisory layer eating a budget. The supervision was supposed to be the cheap part. Check the work, approve or reject, move on. Instead

it became the workload.

39. Lenar Kess 00:10:53

    On the same day, from a different direction, someone running a legal research database posted what they call Rule Zero on the Claude subreddit. The rule is: before an agent is allowed to declare success, it must compare its result against an external check. Not its own assessment of whether it did the job — something outside itself.

40. Damra Vol 00:11:14

    Which reads as obvious for about four seconds and then stops being obvious. Because the expensive question is what counts as external. A test suite the same agent wrote isn't external. A second model asked whether the first model did well isn't external either — that's the six-thousand-turn machine with a nicer name on it.

41. Lenar Kess 00:11:35

    So external means something with independent authority. The database returns the row or it doesn't, the document exists or it doesn't, and the build either compiles or fails.

42. Damra Vol 00:11:46

    And in legal research that's tractable, because a citation either resolves to a real case or it doesn't. Their domain hands them a ground truth for free. Most domains don't, and the reason Rule Zero is hard to adopt generally is that most people can't name their external check without doing real design work first.

43. Lenar Kess 00:12:05

    Then Ethan Mollick posted the third version of this from the user's chair, and it's funnier than the other two. He's telling Sol inside Codex that he'd like to speak to the manager — asking the top-level model to stop delegating his work to dumber agents and elaborate test harnesses that miss things.

44. Damra Vol 00:12:23

    [laugh] Asking to speak to the manager. That's the whole complaint about orchestration in seven words. He paid for the good model and got routed to a call center it built for itself.

45. Lenar Kess 00:12:34

    And notice what all three people are describing. Faxan found a supervisor that wouldn't stop. Rule Zero exists because supervisors mark their own homework. Mollick wants less supervision and more of what he paid for. Same layer, three different complaints.

46. Damra Vol 00:12:50

    There's a fourth item circling the same drain — a Show-HN-grade post about a Model Context Protocol interceptor that sits in front of an agent's tool calls and blocks reads of environment files and dangerous commands in real time. Twelve points and two comments, so I'd call it an example rather than a recommendation. But the existence of a category called interceptor is the interesting signal.

47. Lenar Kess 00:13:14

    Because it means people have concluded they can't get the behavior they want by asking politely in a prompt.

48. Damra Vol 00:13:20

    And Viksit Gaur posted the architectural version of the same frustration — that there's no good primitive for chaining work into a durable workstream, so everyone builds a supervisor loop by hand and every hand-built supervisor loop has the same bug. That connects back to the managed-runtime pitch. Durable execution is on the list of things they want to sell you because everybody's version of it is broken.

49. Lenar Kess 00:13:45

    Which is the one connection across today I'll make. The vendors are selling the layer that three separate people spent yesterday complaining about.

50. Lenar Kess 00:13:53

    We covered the DeepMind leadership transition on Thursday as a straight succession — Hassabis handing the chief executive role to Koray Kavukcuoglu. There's a new claim circulating about why, and let me say up front: this is unconfirmed. It's a post on the singularity subreddit relaying a claim from Chubby, sourced to what they call pathfounders. No named reporting behind it.

51. Damra Vol 00:14:17

    The claim being that Hassabis wanted to leave alongside Jeff Dean, and was talked out of it because Google was worried about what the stock would do.

52. Lenar Kess 00:14:25

    If that's true — and it's a big if — it's a different story than the one we told on Thursday. It would mean a frontier lab's leadership structure was set by equity-market optics rather than by anything about the research.

53. Damra Vol 00:14:39

    And I'd note it's not even an unflattering story about Google. It's just an unusually visible instance of something normal. Retention packages exist. Boards worry about founder departures. What makes this one different is that the founder in question is the one whose name people attach to the science.

54. Lenar Kess 00:14:57

    There's a competing account in the same subreddit. That one says the handoff was voluntary — he considers the research problem close enough to solved that infrastructure is the harder remaining piece, and he'd rather do the science.

55. Damra Vol 00:15:10

    Those two stories aren't mutually exclusive, which is why I think the speculation is going to run for a while. You can want to leave the chief-executive job, be talked into keeping a title, and also believe the interesting work has moved. All three at once is an ordinary way for a career to go.

56. Lenar Kess 00:15:27

    The third item in the same neighborhood is a talk where he says he expects all diseases to be cured within twenty years.

57. Damra Vol 00:15:34

    [breath] Which is a prediction, not a result, and I'd rather not let it carry any weight. He's said versions of this before, and twenty years isn't the number to argue about. What he thinks the bottleneck is now — that's the claim with content in it. If his answer really is infrastructure rather than modeling, then Isomorphic Labs is where that belief gets tested, not DeepMind.

58. Lenar Kess 00:15:56

    That's what I'd hold onto from the whole cluster — the reported reason is unverified, but the claim about where the difficulty has moved is one he's making in public under his own name.

59. Lenar Kess 00:16:06

    SAP has stopped most travel and most hiring, and the reason given internally is AI spending. This comes from 404 Media, working from an internal email, and Bloomberg had the story first in July. One clarification on timing: the freeze went in last month and is still in effect. It reached a wider audience today because it hit Hacker News rather than because anything new occurred.

60. Damra Vol 00:16:31

And the exception is what makes it legible. AI-related travel and AI-related hiring are exempt. So this is a reallocation with a direction rather than a general downturn dressed in AI language. The internal line is that they need to be disciplined in how they spend, while rolling out a newly built AI tool across the whole company.

61. Lenar Kess 00:16:52

    A current employee told 404 Media that the rollout massively increases the costs. Which is a remarkable sentence to have inside the same company as the hiring freeze.

62. Damra Vol 00:17:02

    It's the first hard instance I've seen of a token bill competing directly with headcount at a company that size. We talked about enterprise token spend in passing yesterday as a trend. This is the trend expressed as a decision that changes whether specific people get hired.

63. Lenar Kess 00:17:18

    Let me be fair to SAP here, because there's a reading where this is good management. They found a new cost line that's growing fast, they slowed two discretionary lines to fund it, and they told people why. That's more transparent than most companies manage.

64. Damra Vol 00:17:33

    It's defensible. What I'd note is which lines got chosen. Travel and hiring are the two levers you pull when you want the number to move this quarter without touching anything structural. Nobody renegotiated a model contract, and as far as we know nobody moved a workload to a cheaper tier. They cut the things that were easy to cut.

65. Lenar Kess 00:17:53

    On the supply side there's a related read circulating on the OpenAI subreddit — someone charting OpenRouter usage and arguing that OpenAI's Sol and Luna split is doing what it was designed to do. The heavy reasoning model takes the hard work, and the cheap high-volume model takes everything else.

66. Damra Vol 00:18:11

    That's a screenshot of third-party routing data with someone's interpretation attached, not OpenAI's numbers, so I'd hold it loosely. But directionally it fits what SAP's problem implies. The vendor answer to cost pressure was a cheaper tier rather than a cheaper flagship. The flagship stays expensive and you're invited to use it less.

67. Lenar Kess 00:18:33

Which puts the routing decision on the customer. You now have to know which of your requests deserve the good model, and most organizations have no idea.

68. Damra Vol 00:18:42

    And that's an engineering problem that looks like a procurement problem, which is the kind that sits unowned for eighteen months. Somebody at SAP is building a spreadsheet about it right now.

69. Lenar Kess 00:18:53

    Miles Brundage posted a run of arguments across yesterday. Start with his answer to what the most mistaken idea in AI currently is. He says it's the belief that companies already have the right incentives and the right laws pointing them in a good direction. Coming from someone who ran policy inside OpenAI, that's a specific claim, not a general mood.

70. Damra Vol 00:19:14

    And he sharpens it into a trade-off I hadn't seen put that way. His argument is that the path is narrow between two failures. One is a self-regulatory body for AI along the lines of FINRA in finance, but with no actual safety floor underneath it. The other is rules that exist on paper and get applied selectively, by whoever currently has access.

71. Lenar Kess 00:19:37

    So the two ways to fail are a body with no teeth, and teeth that only bite the people out of favor.

72. Damra Vol 00:19:44

    And they fail in opposite directions, which is why the path between them is narrow. A floor with no enforcer is theater. An enforcer with no floor is leverage. You need both or neither works, and building both at once is much harder than building either.

73. Lenar Kess 00:19:59

    He also points at a writeup of the federal effort from Jay Obernolte and Lori Trahan, and says it improved from a federalism standpoint. Which is a narrow, technical compliment — he's talking about how the bill handles the relationship between federal rules and state ones, not whether he likes the substance.

74. Damra Vol 00:20:18

    That distinction matters more than it sounds. Most of the fight over federal AI legislation in the last year has been about preemption — whether a federal bill wipes out state laws that already exist. If

the current draft handles that better, that's a real change in what the bill would do, separate from what it says it wants.

75. Lenar Kess 00:20:37

    He also flagged an industry standards effort, guidelight dot AI, in the same run of posts. And Micah Carroll added a jab from a different angle — aimed at people who spent years arguing alignment would be easy if anyone made it a priority. Carroll's line is roughly, well, this is their moment to demonstrate that.

76. Damra Vol 00:20:56

    [chuckle] Which is a fair shot, and also a slightly cheap one, because the people who said that mostly meant it as an argument about resource allocation rather than a promise. Still, the incidents of the last two weeks did move a lot of people who were previously comfortable.

77. Lenar Kess 00:21:12

    I'll keep the weight proportionate here — these are tweets, not a paper, and Brundage is arguing rather than reporting. But the reason I'd read him is that he names a specific bill and a specific mechanism instead of asking for regulation in general, which is what most of this discourse settles for.

78. Damra Vol 00:21:30

    Owain Evans pointed at some Apollo Research reading in the same window, if you want the technical version of the same worry rather than the institutional one.

79. Lenar Kess 00:21:39

    Four short items to close, and they're all better than their headlines. Yesterday we talked about DeepSeek V4 Flash's benchmark numbers as vendor-reported, and the gap we named was that DeepSeek's 82.7 percent on Terminal-Bench 2.1 came out of something they call a minimal-mode harness that hasn't been released. Someone has now run it through a public harness across 445 trials, and got the same number back.

80. Damra Vol 00:22:06

    With one disclosure that matters: the person who ran it is the author of the harness they ran it on. That doesn't make it wrong, and 445 trials is a serious number — most benchmark claims you see are single-digit runs. But the independence is partial, and they said so themselves, which I'd rather have than the alternative.

81. Lenar Kess 00:22:27

So the number survived contact with a different harness. That was the open question and now it's mostly closed.

82. Damra Vol 00:22:33

    The second measurement is my favorite thing all weekend, and it's someone noticing a tokenizer difference. They fed the same 330 lines of HTML and JavaScript to two local models. Qwen's 35 billion parameter model turned it into 1,609 tokens. Gemma's 26 billion parameter model needed 4,258 for the identical input.

83. Lenar Kess 00:22:58

    Two and a half times the tokens for the same file.

84. Damra Vol 00:23:01

    For the same file. And that's a cost multiplier and a context multiplier hiding inside a design choice nobody advertises on a model card. Your context window is two and a half times smaller in practice on one of them, for code specifically. It also explains a difference people have been describing by feel for months without a number attached to it.

85. Lenar Kess 00:23:22

    Caveat being one person, one sample, and one file type.

86. Damra Vol 00:23:26

    One sample, yes. I'd still go run it on my own code today, because if it replicates it changes which model you pick for a coding task for reasons that have nothing to do with how smart either model is.

87. Lenar Kess 00:23:39

    Third, and staying with local inference: someone spent a few months writing an inference engine in plain C99 to run 1.58-bit BitNet models. It doesn't use Python, CUDA, or a linear algebra library — just a compiler and a makefile. They report 36 tokens per second on a Xeon.

88. Damra Vol 00:23:59

    And the writeup is mostly about where the memory-bandwidth ceiling starts to bite, which is where it earns its length. Ternary weights make the arithmetic almost free, and then you discover the arithmetic was never the constraint — moving the weights was. That lesson generalizes well beyond BitNet.

89. Lenar Kess 00:24:16

Last item, and it's a small one. Ars Technica reports that ChatGPT has started refusing direct requests to write in a named author's style. The URL carries a July date, so I'd treat it as recent rather than brand new. The piece notes the model can still get somewhere near the same feeling by other routes.

90. Damra Vol 00:24:36

    So it blocks the phrasing rather than the output, which is a policy about how you ask. Over on the Hacker News thread, the top comment is unimpressed on capability grounds — the argument being these models can't write like an average person convincingly, let alone a famous one. I don't agree with all of that, but it's a reasonable objection to the premise that this was ever a strong capability worth restricting.

91. Lenar Kess 00:25:00

    What makes it notable to me is that OpenAI did it without a ruling forcing them to. No court, no settlement, just a product decision on a surface that hundreds of millions of people touch.

92. Damra Vol 00:25:12

    Which puts them in the position of deciding what a style is. That's a question that has occupied copyright law for about a century without resolution, now being answered by a refusal classifier.

93. Lenar Kess 00:25:23

    One gap stays open from the first segment. Nobody outside Anthropic and LangChain has published real numbers on what the managed runtimes cost under load, and every argument about renting the layer turns on that price. Faxan's twenty-five to one ratio is what happens when you can't see it, and renting doesn't make it visible. It just moves whose graph it shows up on. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic