DISPATCH 067 · 2026-06-24 | GSV THE ACCESS PATH HAD A PAGER

When Access Became Part of the System

2026-06-24 / 00:24:01

> *"A model can be good enough to test classified systems and still be the wrong dependency if the access path can vanish underneath the operator."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Today's Braid follows a pressure test: once AI systems enter sensitive work, the vendor relationship, the permission model, the data path, and the audit trail become part of the technology itself.

- Techmeme's pointer to the Mythos 5 NSA reporting anchors the lead: parts of the NSA reportedly red-teamed Anthropic's model for classified-system cybersecurity work, then lost access during the Anthropic dispute.
- The Financial Times report on banned Nvidia chips in China turns export control from policy language into prices, scarcity, and workarounds.
- The Linux Foundation's Agent Name Service announcement gives the agent segment a concrete artifact: DNS-backed naming, verification, and discovery for internet-facing agents.
- SAFARI, LemonHarness, and OpenThoughts-Agent show the research side of the same work: agents need fault attribution, bounded workspaces, time awareness, and better training data.

- The Guardian's Schneier and Sanders essay and its Meta employee-tracking report move accountability into two places builders can inspect: AI output liability and worker-data consent.
- The open-source coding-agent census and the GUI-versus-CLI benchmark close on measurement: agent adoption and agent execution are now being studied through traces, interfaces, and verifiers instead of vibes.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Techmeme's pointer to the New York Times and follow-on government-tech reporting says parts of the NSA were red-teaming Anthropic's Mythos 5 against classified-system cybersecurity work, and then some of that access went away during the Anthropic dispute. [pause] That's the lead today because the story starts with a capability test and ends with a dependency test. If you're the agency, you aren't only asking whether the model can identify flaws. You're asking who can turn it off, under what order, and whether your sensitive workflow survives the loss of that access.

   - techmeme.com
   - nextgov.com
   - ft.com
   - reddit.com
   - axios.com

- techmeme.com
- techmeme.com
- linuxfoundation.org
- arxiv.org
- arxiv.org
- arxiv.org
- theguardian.com
- theguardian.com
- spectrum.ieee.org
- techmeme.com
- techmeme.com
- forbes.com
- arxiv.org
- arxiv.org
- reddit.com
- reddit.com

2. Damra Vol 00:00:38

   And the strange part is that those questions arrive after the model has already crossed the first trust threshold. The agency wasn't reading a launch post. It was testing a frontier system against sensitive machines. Once you do that, model access stops being a procurement detail. It's part of the operating plan.

3. Lenar Kess 00:00:58

   Right. The reports I could verify are secondhand in a few places, so I'm not going to overclaim what Mythos found inside those systems. The sourced version is narrower and still enough: a government customer was evaluating Anthropic's new model for defensive cyber work, the Trump administration's export-control action forced tighter access limits, and parts of the NSA lost the easy path to the tool. Business Insider also reports a related lawsuit from Legion AI over losing access to Anthropic's Fable 5 and Mythos 5, which gives the same story a commercial face.

4. Damra Vol 00:01:36

   That commercial face matters because it keeps this from being a spooky government anecdote. A startup can write a contract around an API. A federal agency can write a program around a model. Both can still discover that the policy layer above the vendor is now part of the availability model.

5. Lenar Kess 00:01:53

And that has a builder-facing consequence. A model involved in red-team work, incident response, or classified review needs more than a backup model name in a config file. The failover plan has to cover nationality checks and credential revocation. It also needs audit records, data retention, and a way to replay the work without the original provider.

6. Damra Vol 00:02:15

   The replay problem keeps nagging at me. If Mythos finds a flaw and the agency loses access the next day, can the agency preserve enough of the trace to patch the flaw and prove the model didn't touch material it shouldn't have touched? Can it hand the case to a human analyst? If the answer depends on a vendor console you can't open anymore, the artifact was never fully yours.

7. Lenar Kess 00:02:37

   That also explains why this isn't a rerun of the agentic-cyber stories from earlier this week. We already talked about offensive capability and security plugins. Today's version is more institutional. The model may be capable, but the institution has to absorb access churn, legal orders, and vendor disputes without losing the audit trail for the work it already did.

8. Damra Vol 00:03:01

   Anthropic shouldn't carry the whole story alone. The company is responding to a government order, not waking up and deciding to annoy its customers for sport. But for anyone building on top of a frontier model, motive doesn't change the failure you have to design around: a dependency can disappear for reasons that have nothing to do with uptime.

9. Lenar Kess 00:03:21

   Exactly. A builder should treat access as something the system tests. Accuracy isn't enough. Test revocation and degraded mode. Test artifact export and audit replay. Then test whether the team can keep working when the vendor, the regulator, and the customer's compliance office all have a say.

10. Lenar Kess 00:03:41

    The Financial Times reports that banned Nvidia AI chips have more than doubled in price on China's black market. The DGX B300 server is reported above eight million yuan, roughly 1.1 million dollars, up from around four million yuan over six months; the RTX 6000 Pro workstation chip is reported up from fifty thousand to one hundred thirty thousand yuan.

11. Damra Vol 00:04:06

That turns export control into something you can see on an invoice, even if the invoice isn't exactly the one regulators wanted to exist. [tsk] The policy lever is access to accelerators. The market response is scarcity pricing, smuggling risk, substitute hardware, and domestic-chip storytelling.

12. Lenar Kess 00:04:25

    I like this as a follow-up to yesterday because it is narrower than yesterday's chip-diplomacy story. When enforcement tightens, routing around the restriction gets more expensive. Techmeme's chip item points to the FT report; the LocalLLaMA post maps seven Chinese companies claiming H100- or H200-class chips; Axios covers the GLM-5.2 debate around Chinese frontier models; and Techmeme has Zhipu capital-market heat plus Qualcomm-ByteDance custom-chip talks.

13. Damra Vol 00:05:00

    I'd keep the Reddit chip map at arm's length until the company list is checked one by one. It may be operator color, but it's not the same kind of source as a shipment record or a teardown. The stronger claim is already there without it: the restriction is producing a visible premium, and Chinese buyers are still trying to get compute through whatever channel works.

14. Lenar Kess 00:05:22

    Yes. And the model side matters too. Axios puts GLM-5.2 in the middle of the China-versus-US conversation, after a week where builders were already testing how close open and Chinese systems are getting to Western frontier models on agentic work. I don't think we need to make this mystical. If the chip channel is constrained but not closed, and if local models keep improving, the next few months become a lot of messy substitution rather than one neat policy outcome.

15. Damra Vol 00:05:53

    That substitution can still be expensive and worse. A black-market server with no normal support path isn't a data-center strategy. Domestic silicon that looks good on paper may still have tooling gaps, memory limits, compiler pain, and weird failure cases. Blocked buyers often pay more, accept more friction, and keep moving anyway.

16. Lenar Kess 00:06:15

    That gets us to capital. Zhipu reportedly drawing market heat and ByteDance exploring custom silicon with Qualcomm don't prove China has solved accelerator dependence. They show the industry treating restricted compute as a business problem, not only a foreign-policy problem. The engineering question underneath is how much performance you can recover through model architecture, software, batching, distillation, and less-preferred chips before the missing Nvidia path hurts too much.

17. Damra Vol 00:06:44

    That question is harder than the public argument usually admits. A model lab can make a smaller or more efficient model look impressive. Serving it reliably and cheaply, with good developer tooling and enough memory bandwidth, is a different exercise. Export controls push on the hardware input, but the response spreads into compilers, interconnects, cloud rental pricing, procurement, and model design.

18. Lenar Kess 00:07:08

    So the China segment isn't 'controls failed' or 'controls worked.' We don't have that evidence. What we have today is a price signal, a substitution signal, and a capital signal. If those keep lining up, the practical world builders deal with will be less like a denied market and more like a more expensive, more fragmented, less supportable stack.

19. Lenar Kess 00:07:31

    The Linux Foundation announced an intent to launch Agent Name Service, or ANS, as an open standard for trusted identity, verification, and discovery of AI agents using existing Domain Name System infrastructure. Their own announcement says the idea is to let organizations identify agents through domain names they already own, without a proprietary namespace or gatekeeper.

20. Damra Vol 00:07:55

    That is exactly the kind of standard that sounds underwhelming until you try to delegate work across systems. If an agent can call another agent, fetch a tool, talk to a Model Context Protocol server, or act for a company, you need to know which entity you are talking to and who vouches for it. DNS isn't glamorous, but it is already how the internet answers a lot of identity-adjacent questions.

21. Lenar Kess 00:08:20

    The timing is good because the research side today is full of agent control work. SAFARI, the arXiv paper on long-horizon fault attribution, argues that loading an entire multi-agent trace into a large language model's context window starts to fail when traces run into millions of tokens. Its answer is an active investigator: read specific trace windows, search for patterns, keep a short-term memory, and verify atomic claims before naming the decisive fault.

22. Damra Vol 00:08:51

    That paper has a nice debugging smell. The agent doesn't pretend the whole log fits in its head. It scrolls, searches, keeps notes, and asks evaluators to check the claims against quoted evidence. That

is much closer to how a good engineer investigates a broken run. You don't paste the entire world into a prompt and hope attention saves you.

23. Lenar Kess 00:09:13

    The numbers are worth saying once. SAFARI reports a twenty percent improvement on the Who and When benchmark with a one-million-token budget. It also reports a nineteen percent strict-precision improvement on the TRAIL GAIA subset with a twenty-five-thousand-token budget. In the stress case, it maintains 0.58 precision when the target fault sits five times beyond the model's native context window. Those are paper numbers, not deployment guarantees, but the mechanism is the interesting part.

24. Damra Vol 00:09:45

    And LemonHarness attacks the neighboring problem: not where the failure happened in a giant trace, but how to keep a long-running agent from losing track of the workspace. The paper says mutating actions are a small slice of steps, around fourteen to eighteen percent in the cited SABER result, but one wrong mutating deviation can hurt task success far more than a non-mutating one. File writes, dependency installs, temp artifacts, and background processes need a bounded place to happen.

25. Lenar Kess 00:10:17

    LemonHarness wraps that in a controlled workspace and structured tools. It also gives the agent reusable rule knowledge, execution logs, and time-aware execution. On Terminal-Bench 2.0, the paper reports 84.49 percent accuracy over 445 trials with GPT-5.3-Codex, and 86.52 percent average accuracy across five jobs with GPT-5.5. Again, the exact benchmark result is less durable than the design lesson: the harness is no longer a wrapper around the model. It is part of the model's effective behavior.

26. Damra Vol 00:10:56

    OT-Agent adds the training-data side. The OpenThoughts-Agent paper says it ran more than one hundred ablations, built a one-hundred-thousand-example training set, and fine-tuned Qwen3-32B to 44.8 percent average accuracy across seven agentic benchmarks, a 3.9-point gain over Nemotron-Terminal-32B in their comparison. The detail I like is that the strongest teacher wasn't simply the strongest model by benchmark. Data recipes have taste baked into them.

27. Lenar Kess 00:11:33

    So this chapter has four pieces that belong together without forcing them into one grand claim. ANS says agents need names and verifiable identities. SAFARI says long agent traces need active fault investigation. LemonHarness says state-changing work needs a bounded runtime and time

awareness. OT-Agent says training broadly capable agents depends on the task sources, teachers, filters, and traces you choose. That is a practical stack, not a demo reel.

28. Damra Vol 00:12:05

    Google DeepMind's public multi-agent discussion fits as a broad marker rather than a source for a specific technical claim here. Its listing frames the episode around agentic tools and multi-agent systems. The artifacts we did fetch give that idea teeth: identity, memory, execution boundaries, fault attribution, and training data. Those plain nouns decide whether delegation survives contact with a real environment.

29. Lenar Kess 00:12:33

    I'd phrase it this way: a useful agent is less like a clever chat session and more like a small institution. It has a name, a credential, a workspace, a memory, a budget, logs, tests, and someone who can say why it failed. If one of those is missing, the agent may still impress you in a demo, but it will make the operator pay later.

30. Lenar Kess 00:12:56

    The Guardian ran a Bruce Schneier and Nathan Sanders essay about a German court ruling that held Google responsible for inaccurate AI-generated search summaries. The reported ruling treated AI Overviews as Google's own content rather than ordinary search links, and it rejected the idea that users should simply verify the generated answer themselves.

31. Damra Vol 00:13:19

    That is a sharp line. Search links point away from the company. An AI summary speaks in the product's voice. If the product invents a connection between two publishers and shady businesses, the court's view is that Google can't hide behind the user needing to be skeptical. The generated answer is part of Google's business activity.

32. Lenar Kess 00:13:38

    Schneier and Sanders connect that to a broader accountability claim: when a company deploys an AI system as part of its service, the system should be treated more like a representative of the company than like a random stranger posting on the internet. They also point to the Air Canada chatbot case, where the airline was held responsible for what its chatbot told a customer.

33. Damra Vol 00:14:00

    That is where the engineering work gets less abstract. If the company is liable for the answer, then the product team needs provenance, escalation, confidence thresholds, refusal behavior, and a

correction path. You can still use generative output, but you can't pretend it is vapor when the customer gets hurt by it.

34. Lenar Kess 00:14:19

    The second Guardian item brings the same accountability problem inside the workplace. Meta paused its Model Capability Initiative after privacy concerns and internal backlash. The reporting says the tool tracked employee keystrokes, mouse clicks, and screen content for AI training; more than sixteen hundred workers signed a petition; and Meta paused the program after a Wired-reported internal security issue exposed potentially sensitive collected data inside the company.

35. Damra Vol 00:14:48

    That is both a consent problem and an access-control problem. Employees were already objecting to the premise: my work machine isn't raw material for training just because my employer owns the laptop. Then the security issue made the internal-data question concrete. Who could see the captured data? Was sensitive content separated? Could performance information, private messages, or unfinished work leak across teams?

36. Lenar Kess 00:15:14

    And again, motive isn't the whole story. Meta can argue that employee activity is useful for training coding and intelligence skills. It probably is. A company with thousands of engineers sitting inside its systems has a tempting corpus of work traces. But useful data isn't automatically legitimate data, and high-value internal data tends to attract bad access patterns unless the boundaries are designed before collection begins.

37. Damra Vol 00:15:41

    The builder version is uncomfortable because every team wants better traces. You want to know how experts use the tool, where they hesitate, what shortcuts they take, and what context they keep open. But if your collection path records everything first and sorts out consent later, you have already made the trust decision. The pause doesn't erase the fact that the system existed.

38. Lenar Kess 00:16:05

    So the court story and the Meta story aren't the same. One is user-facing output liability. One is worker-data collection. But they both push AI deployment back into ordinary institutional responsibility. If the company benefits from the system, it has to own the answer, the data path, and the people exposed by both.

39. Lenar Kess 00:16:25

IEEE Spectrum's FERC piece says US regulators are trying to speed data-center grid connections without pushing the cost onto ordinary customers. That is the most direct way to mention today's power story without spending another full episode on energy. The data-center queue is now a policy object, and regulators are trying to decide who pays when giant new loads want priority.

40. Damra Vol 00:16:49

    And this is where the word 'compute' hides too much. A data center isn't only GPUs. It is interconnection studies, substations, transmission upgrades, backup generation, water, and a local rate base that may not want to subsidize a hyperscaler. The FERC issue isn't whether AI is exciting. It is whether the grid process can distinguish a serious load from a speculative land grab and allocate costs in a way that survives politics.

41. Lenar Kess 00:17:18

    The market side is just as heavy. Techmeme points to an AWS CEO interview discussing a planned two hundred billion dollars of 2026 capex and new agentic capabilities in recruiting and coding. Techmeme also has SK Hynix seeking major US capital for capacity, and Forbes covers memory-chip suppliers being repriced by the AI data-center buildout.

42. Damra Vol 00:17:42

    Memory deserves its own sentence because it keeps becoming the constraint people remember late. Training and serving models need accelerators, but high-bandwidth memory supply decides how many of those accelerators are useful and at what margin. If SK Hynix is raising for capacity and memory suppliers are being valued like AI infrastructure companies, the market is reading the stack below the model.

43. Lenar Kess 00:18:07

    I want to keep this proportionate because we've been on compute for days. Today's infrastructure note isn't a new turning point. It is another set of receipts: grid policy, hyperscaler capex, memory capacity, and export-control pricing are all showing up in public numbers. The model news is easier to read, but the capacity news is where a lot of the future bill arrives.

44. Damra Vol 00:18:31

    And the bill changes behavior before the capacity arrives. Teams pick smaller models. Labs chase better inference efficiency. Cloud buyers lock in supply. Governments ask whether the power deal is a private business cost or a regional development project. None of that has the neatness of a benchmark table, but it is the work that decides which benchmark table can be served to users.

45. Lenar Kess 00:18:54

One arXiv paper today tries to measure coding-agent traces across more than 180 million Git repositories using World of Code. The authors combine configuration-file scanning, commit-message analysis, author-identity patterns, and bot-signature lookup. Their main warning is simple: no single signal captures the whole population.

46. Damra Vol 00:19:18

    That paper is a nice antidote to lazy adoption numbers. It says bot-account lookup found 28,154 Claude Code commits in one snapshot, but the multi-method union found 850,157. That is a thirty-times relative-recall gap. If you only look for the visible bot account, you miss the way people use the tool under their own identity or through message trailers.

47. Lenar Kess 00:19:45

    The paper also says PR-based and commit-based views see different worlds. Codex appears as the largest agent by pull requests in the AIDev comparison, while Claude Code is near absent there and large in commits. Claude Code is large in commits and much smaller in pull requests. That means a measurement can accidentally become a product-interface study. You count the channel the tool leaves behind, not a pure concept called agent adoption.

48. Damra Vol 00:20:15

    That should make every headline about coding agents name its detector. A cloud agent that opens pull requests, an editor agent that commits locally, and a terminal agent whose trace is buried under a human author name aren't equally visible. The census is still valuable, but the detector is part of the claim.

49. Lenar Kess 00:20:35

    The GUI-versus-CLI benchmark hits the same theme from the execution side. It compares 440 desktop tasks across 18 applications with matched goals, initial states, and final-state verifiers. The strongest GUI agent, GPT-5.4, gets a 59.1 percent full pass rate. Under the original CLI-Anything skill layer, Codex GPT-5.5 is the strongest original skill-mediated CLI agent at 48.2 percent.

50. Damra Vol 00:21:09

    Then the diagnostic result changes the conversation. Only 37.6 percent of verifier checkpoints were satisfiable by the original skill interface. When the researchers patched the skills using verifier-observed requirements, CLI success rose to 69.3 percent. They correctly call that an upper bound, not a fair deployed baseline, because the repair used verifier information. But it shows how much performance can live in the interface instead of the model.

51. Lenar Kess 00:21:39

That is the most practical builder point in the measurement segment. If an agent fails through a GUI, maybe it missed the button, misread the screen, or gave up early. If it fails through a CLI skill layer, maybe the model did everything the skill exposed and the missing operation was never available. Your eval needs to distinguish model failure from interface failure, or you will improve the wrong thing.

52. Damra Vol 00:22:04

    The LocalLLaMA Qwen-AgentWorld item fits as a smaller note here. A 35-billion-parameter mixture-of-experts model with three billion active parameters is being discussed as a simulator for Model Context Protocol, terminal, software-engineering, Android, web, and operating-system environments. I wouldn't make too much of a Reddit post alone, but the direction makes sense: if agents need training data, simulated environments become part of the toolchain.

53. Lenar Kess 00:22:40

    And the ClaudeAI post about software entering an 'infinite monkeys' era is useful mostly as a mood check. More people can generate code, which means the scarce skill moves toward judging, integrating, testing, and maintaining the code. That isn't a new sermon for this show. The evidence keeps pointing at the same review loop: traces, verifiers, skills, workspace boundaries, and accountability records.

54. Damra Vol 00:23:07

    The better version of that anxiety isn't contempt for new builders. It is care for the review loop. If a beginner can create a pull request with a capable agent, great. The ecosystem still needs detectors that know where agent code appears, interfaces that expose the right operations, and maintainers who can tell a useful change from a plausible one.

55. Lenar Kess 00:23:28

    That brings us back to the NSA story without forcing a bow onto it. Today's sources kept asking who owns the work after the model acts: the agency that lost access, the buyer paying gray-market prices, the company naming its agents, the court assigning liability, the employer collecting work traces, and the researcher measuring commits. The record that survives after the system has finished its turn is going to decide how much of that work people can trust. Lenar.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic