Who Gets to Approve the Model

2026-06-18 / 00:22:51

> *"If model access is decided in private calls, your product still needs a public answer for what happens when that access changes."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Today's episode follows a practical tension in AI: the public story says less regulation, while the daily operating reality is access reviews, model blocks, cloud dependencies, water constraints, and private data moving through agents.

- Axios on the White House AI power center maps the officials now shaping AI decisions, from Howard Lutnick to Scott Bessent and Ryan Baasch.
- Axios on Trump's shadow AI policy explains why an anti-regulation posture can still produce case-by-case intervention.
- Techmeme's Wired summary on Anthropic and SK Telecom shows how national-security pressure turns into a specific model-access decision.
- Axios on Arizona data centers grounds the compute story in power, water, grid costs, and local utility planning.
- Rest of World on Chile's undersea cable fight shows network routes becoming geopolitical infrastructure, not neutral background.

- SafeClawBench, TRAP, and the agent-memory paper turn the research block into one builder question: what state, evidence, and private data should survive contact with tools?

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Axios has two pieces this morning on how AI decisions are getting made inside the Trump White House. One is about the people: Howard Lutnick gaining influence after the Anthropic fight, Scott Bessent showing up in industry discussions, Susie Wiles and Ryan Baasch becoming more important, and the earlier Silicon Valley center around David Sacks and Sriram Krishnan giving way to a wider set of officials. The other piece says the administration came in opposed to formal AI regulation, but is still shaping the industry through case-by-case interventions.

   - axios.com
   - axios.com
   - techmeme.com
   - x.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - techmeme.com
   - restofworld.org

- axios.com
- cnbc.com
- techmeme.com
- arxiv.org
- arxiv.org
- arxiv.org
- arxiv.org
- techmeme.com

2. Damra Vol 00:00:38

   That distinction matters because it changes the question for builders. A formal rule may be annoying, but you can read it, model it, and test against it. Case-by-case pressure is harder. You can wake up with a vendor, a cloud path, or a regional user group suddenly sitting inside a policy argument you didn't know you were part of.

3. Lenar Kess 00:00:58

   Exactly. I don't want to replay the Anthropic export-control story we already covered this week. The new development today is the operating model Axios is describing. Publicly, the posture is less regulation, less of the Biden-era rule structure, and more room for companies to move. In practice, Axios points to state-law preemption and federal procurement. It also points to national-security review and direct pressure around specific model releases or specific partners.

4. Damra Vol 00:01:29

   The Anthropic example helps because it is concrete. Techmeme's Wired summary says the White House move to restrict Mythos 5 came after it ordered Anthropic to revoke SK Telecom's access over alleged China ties. That isn't a generic policy mood. That is a named model, a named company, a named customer, and an access decision.

5. Lenar Kess 00:01:52

   Right, and Gabby Miller also flagged the legislative version of the same instinct: Representative Josh Gottheimer wants top AI companies to submit powerful models to the government, while the White House path has been more voluntary-review language and direct executive pressure. Those aren't identical mechanisms. One is proposed legislation. One is an executive operating style. But both point at the same practical boundary: frontier model access is no longer only a contract between a lab and a customer.

6. Damra Vol 00:02:23

[tsk] Voluntary doesn't buy you much comfort there. If the federal customer, the export-control office, the procurement path, and the national-security review all sit on one side of the table, the lab can call participation voluntary and still feel the cost of refusal.

7. Lenar Kess 00:02:40

   My read gets sharper at the architecture boundary. If you build on these models, you can still prefer the best model. I do. The craft point isn't that everyone should run away from frontier APIs. It is that access risk now belongs in the architecture. You need to know what breaks if a model becomes unavailable in one region, for one customer type, or under one government condition.

8. Damra Vol 00:03:03

   And you need the answer before the policy event. A fallback that only exists after the model gets blocked is a press statement, not an engineering path. For a serious product, the maintenance work stops feeling optional as soon as customers depend on it. Model routing, degraded modes, data retention, and entitlement checks become part of the product claim.

9. Lenar Kess 00:03:26

   I would put it this way: the approval surface moved. It used to be mostly internal, meaning procurement, legal, security review, and maybe a data-processing agreement. Now there is a second approval surface outside the company, and it may not publish a spec. Axios is useful today because it names the people and the pattern, but the engineering consequence is pretty plain. Your dependency graph now includes offices you don't control.

10. Damra Vol 00:03:53

    There is also a fairness problem under that. The companies with the best government relationships can navigate a phone-call policy model better than smaller labs, foreign customers, or downstream startups. I don't have a source saying that happened in this case; that is just how permission systems behave when the criteria are partly social. The people with the best map move faster.

11. Lenar Kess 00:04:16

    And it makes trust harder for foreign buyers. Construct covered that yesterday from the sovereign-access angle, so I won't belabor it. But today adds the domestic machinery: who in Washington has influence, how the administration reconciles anti-regulation language with intervention, and why a company can be told to change access without a statute that looks like Europe's AI Act.

12. Damra Vol 00:04:38

That also explains why this story is bigger than Anthropic, even though Anthropic keeps appearing in the examples. The model provider is the visible party, but the policy action travels through banks, telecoms, federal buyers, and cloud accounts. It shows up where a user actually touches the system.

13. Lenar Kess 00:04:56

    Yes. And that takes us to the next set of stories, because the access question isn't just Washington asking a lab to say yes or no. It is also which cloud bill can exist, which bank employee can open Claude, which supplier gets the AI boom, and which cable route carries the data.

14. Lenar Kess 00:05:15

    Bloomberg, through Techmeme, reports that ByteDance has been Microsoft's biggest AI customer in recent years, largely using OpenAI models, and is on track to spend more than a billion dollars a year on Azure services. Financial Times, also through Techmeme, says JPMorgan Chase stopped staff in Hong Kong from accessing Anthropic's models, after a similar move by Goldman Sachs.

15. Damra Vol 00:05:40

    Those two items belong near each other, but they shouldn't be collapsed into one China story. ByteDance spending heavily on Azure is a customer-and-cloud story. JPMorgan blocking Anthropic access in Hong Kong is a corporate risk-control story. They both touch geopolitics, but the mechanism is different.

16. Lenar Kess 00:06:00

    That is the important distinction. The ByteDance report says a Chinese company can be a very large customer for a U.S. cloud provider and, indirectly, for U.S.-controlled frontier model access. The JPMorgan report says a global bank can decide the risk is too high for a particular tool in a particular market. One side looks like commercial entanglement. The other looks like institutional caution.

17. Damra Vol 00:06:26

    And both make the vendor abstraction leak. A developer sees an API endpoint. A compliance team sees sanctions exposure, data residency, employee location, vendor nationality, and whether the model provider might get a call from Washington. The product is the same, but the operating reality is not.

18. Lenar Kess 00:06:45

    The supply-chain piece is similar. Techmeme's New York Times summary says South Korean and Taiwanese tech companies that helped build China's hardware sector during the smartphone boom

are now benefiting from the AI boom as U.S. curbs sideline China. Again, that isn't a model-card change. It is a redirection of who gets demand when rules constrain one path.

19. Damra Vol 00:07:08

    And it isn't morally neat. A supplier can benefit from restrictions without being the author of the policy. A cloud provider can sell into demand that policy also makes fragile. A bank can block a tool and still be desperate for AI productivity elsewhere. The operator view is messy because the system is messy.

20. Lenar Kess 00:07:27

    Rest of World's Chile cable story gives the most physical version of this. Juan Ortiz-Freuler reports that Chile was assessing a 500 million dollar China Mobile proposal to connect Valparaíso and Hong Kong by undersea cable. The U.S. revoked visas for three Chilean officials, saying their activities compromised critical telecommunications infrastructure and regional security. Google already has the Humboldt cable linking Chile to Australia expected in 2027.

21. Damra Vol 00:07:57

    That story is almost too perfect as a systems diagram. The cable carries latency, redundancy, diplomatic exposure, cloud dependency, and surveillance anxiety in one route across the ocean. Rest of World quotes former Chilean diplomat Jorge Heine saying Chile has tried to maintain broad economic relationships without being forced to choose between major powers. That is a very different sentence when the object is a fiber route.

22. Lenar Kess 00:08:26

    And Pedro Huichalaf, a former undersecretary for telecommunications in Chile, put it in resilience terms: Chile should have a main and secondary route to Asia in case one fails. That is the builder sentence in the piece. Redundancy sounds obvious until the redundant path is owned by the partner a government wants to keep out.

23. Damra Vol 00:08:47

    It also makes the AI access conversation less software-only. If a model call crosses an ocean, the ocean has owners, contracts, chokepoints, and governments watching it. You can make the endpoint look global in the SDK, but the packets still travel through political geography.

24. Lenar Kess 00:09:05

    So the practical sequel to the White House story is this: access is being decided in more layers than the product page admits. A lab can approve a customer, a bank can block a tool, a cloud can carry a

rival's demand, a supplier can win because a different supplier is constrained, and a cable route can become a diplomatic problem. None of those require a grand theory. They just require reading the dependency chain all the way down.

25. Lenar Kess 00:09:32

    Axios uses Arizona today as a test case for AI's water and power problem. The quote that stuck with me is from Kevin Thompson on the Arizona Corporation Commission: utilities took more than a century to build what Arizona may need to double in the next four to five years to meet demand. That is the data-center story with the romance removed.

26. Damra Vol 00:09:54

    It is also a better update than another capital-spending number. We already talked this week about compute economics. Arizona is a siting story. Where does the power come from? Who pays for the grid connection? How much water does the cooling design need? What happens when the place that has cheap land also has drought and heat?

27. Lenar Kess 00:10:15

    Axios reports that federal electricity regulators may propose rules to accelerate data-center grid connections while limiting costs passed to other customers. It also notes Google shifting to air-cooled systems at a new facility near Phoenix, which reduces direct water use but can change the power profile. That is the trade. You don't delete the constraint; you often move it.

28. Damra Vol 00:10:37

    And moving the constraint can be the correct engineering answer. Air cooling in a water-stressed region may be exactly what you want. But then the grid has to absorb the new behavior, and local residents still care about reliability and bills. A data center isn't a cloud in the local permitting meeting. It is a large electrical customer with a cooling plan.

29. Lenar Kess 00:11:01

    CNBC adds the climate-risk side today, reporting on a study about data centers and exposure to climate hazards. I am keeping that at the level CNBC reports it: capacity planning has to include physical risk, not just GPU availability. Heat, storm exposure, water stress, and insurance all become part of the cost model.

30. Damra Vol 00:11:23

    The Verse funding item fits here too. Techmeme's Bloomberg summary says Verse Enterprises, which wants to provide energy-management software for 100 data centers by 2027, raised a 54

million dollar Series B from Nvidia and others. That doesn't solve the grid problem. It does show where capital is trying to put a control layer around it.

31. Lenar Kess 00:11:45

    That is a modest claim, and it is enough. Energy-management software isn't a magic answer to water rights or transmission queues. But if the constraint is partly operational — when workloads run, how loads flex, and how facilities respond to price and grid stress — then software becomes part of the infrastructure stack in a literal sense.

32. Damra Vol 00:12:06

    And that should make AI builders a little less casual about where compute comes from. The model feels virtual until your inference margin depends on a substation upgrade, a local regulator, or a cooling choice made in a desert county. The abstraction is useful, but it isn't free.

33. Lenar Kess 00:12:25

    There is a temptation to make every data-center story into one giant claim about AI exhausting the world. I don't think today's sources require that. They show something more specific and more actionable: capacity is local. The deciding facts can be water, grid interconnection, climate exposure, local tax policy, or whether the facility can change its load when the grid is strained.

34. Damra Vol 00:12:50

    That specificity matters because it keeps the argument from becoming vibes. A facility in Arizona, a facility in Virginia, and a facility in northern Europe may all serve the same model family and still have very different public costs. The cloud product hides that difference from the developer, but the community and the grid don't get the hidden version.

35. Lenar Kess 00:13:12

    And if you are designing systems that will need lots of inference, that is the deeper planning point. Capacity isn't only a token price. It is a location, a cooling design, a power purchase, a regulator, and a community negotiation. The more agentic work we push into always-on systems, the less those details stay in the background.

36. Lenar Kess 00:13:33

    The arXiv feed is overloaded today, so I am not going paper by paper. The research shelf is narrower: what agent state has to remember, how tool harm should be measured, how private document fields leak, and how long-horizon agents should be evaluated.

37. Damra Vol 00:13:50

Good. Because the paper count can trick you into thinking the day is more research-heavy than it is. For builders, the test is whether these papers help us design agents that keep the right state, produce evidence when something goes wrong, and don't expose private fields just because the user asked nicely.

38. Lenar Kess 00:14:10

    The memory paper, "What Must Generalist Agents Remember?", makes a formal version of a very practical point. If two domains look the same at an observation bottleneck but require incompatible optimal actions, a near-optimal agent has to carry domain-relevant information in memory. The paper uses a ForkWorld gridworld where the agent reaches a fork, the up and down actions are swapped depending on a hidden domain, and a memoryless policy is capped at guessing.

39. Damra Vol 00:14:38

    That is a nice toy setup because it strips away the fashionable parts. No fancy tool chain. No social setting. Just a fork where the current observation is insufficient. If the agent doesn't remember what kind of world it is in, it can't choose the correct arm reliably.

40. Lenar Kess 00:14:54

    SafeClawBench then asks a different question: when a tool-using agent fails, what kind of failure did you observe? The paper introduces 600 adversarial tasks across six attack families. They include direct and indirect prompt injection, tool-return injection, memory poisoning, memory extraction, and ambiguity-driven unsafe inference. It separates semantic acceptance, audit-visible harm evidence, and sandbox-observed tool or state harm.

41. Damra Vol 00:15:23

    That separation is where the benchmark helps. If the model says the wrong thing but never touches state, that is bad in one way. If it writes persistent memory, modifies a database, sends a message, or triggers code, that is bad in another way. One attack-success number hides the part an operator needs to debug.

42. Lenar Kess 00:15:43

    The numbers are sobering without needing drama. The paper says that without additional prompt protection, semantic failure rates ranged from 9.0 percent to 44.2 percent across models. It also reports that in a 12,000-row matched analysis, 291 of 347 observed sandbox harms occurred in rows that passed the semantic check. In plain English: a semantic filter can miss tool-state harm.

43. Damra Vol 00:16:13

That last result is the one I would tape to the eval dashboard. If your test only reads the answer text, you may miss the file write, the poisoned memory, or the database mutation. Agent evaluation has to inspect the world after the turn, not just the sentence the model produced.

44. Lenar Kess 00:16:30

    TRAP, the privacy paper, is even more direct. It studies agents in document-heavy workflows where private information is required to complete the task: a passport number for a flight booking, a bank account number for payroll, that kind of thing. The authors argue that a model capable enough to use private information for the task can also be induced to reveal it.

45. Damra Vol 00:16:52

    And that is painfully close to real product design. You can't simply hide every private field from the agent, because then the agent can't do the job. But you also can't let the model see the private field as ordinary text and trust a prompt to make it behave forever.

46. Lenar Kess 00:17:08

    The paper evaluates 22 models and says every model family showed non-trivial leakage. It also says prompt-based defenses reduced leakage at significant cost to task accuracy, and argues that for softmax-based models, prompt-style soft constraints can't jointly deliver high task success with zero leakage probability. Their proposed answer is structural private-field isolation: replace private fields with hash keys before they reach the model.

47. Damra Vol 00:17:39

    That is the agent-architecture lesson. Don't ask the model to be the privacy boundary if the system can remove the secret before the model sees it. Let the model operate on handles, then have deterministic code resolve the handle only at the tool boundary that actually needs it.

48. Lenar Kess 00:17:55

    CEO-Bench is the fourth paper I would only mention briefly. It simulates an agent operating a startup for 500 days through a Python interface. The agent has to handle pricing, marketing, budgeting, noisy data, and long-horizon decisions. The abstract says even strong models struggle, with only Claude Opus 4.8 and GPT-5.5 finishing above the one million dollar starting balance, and neither consistently turning a profit.

49. Damra Vol 00:18:25

    That is a good benchmark if you treat it as a stress test, not a forecast about AI CEOs. Long-horizon agency needs memory, measurement, and privacy boundaries before it needs a title. The papers

today line up around that practical sequence.

50. Lenar Kess 00:18:41

    So the research block isn't a claim that agent reliability got solved on Thursday, June 18. It is a set of test fixtures and arguments. Memory has to carry hidden domain information. Safety evals have to inspect tool effects. Privacy defenses need structure outside the prompt. And long-horizon benchmarks should make the agent live with its earlier decisions long enough for the accounting to catch up.

51. Lenar Kess 00:19:06

    One personnel item before we close: The Information, via Techmeme, reports that Noam Shazeer is leaving Google to join OpenAI as the lead for architecture research. He had rejoined Google as a Gemini co-lead in 2024 through the 2.7 billion dollar Character.AI deal.

52. Damra Vol 00:19:26

    This is hiring news, so it deserves restraint. We shouldn't turn one reported move into a strategy document. But Shazeer isn't random executive churn. His work sits close to the model-architecture layer, and that layer still matters a lot even when the public conversation gets pulled toward products and policy.

53. Lenar Kess 00:19:45

    That is my read too. The industry often talks as if scaling, product distribution, and regulation are the whole game now. But architecture research is still where a lab can change the cost or capability curve in ways that product teams feel later. If OpenAI is bringing Shazeer in to lead that work, it is fair to note the signal without pretending we know the internal plan.

54. Damra Vol 00:20:08

    It also rhymes with the rest of the day in a limited way. Access and compute are getting harder, so efficiency and architecture taste become more valuable. That doesn't mean this hire explains OpenAI's roadmap. It means frontier labs still compete over the people who can make the model itself less wasteful, more capable, or easier to train.

55. Lenar Kess 00:20:30

    It is a reminder not to flatten AI progress into public demos. A lot of the meaningful work happens before the demo exists: the routing choice, the memory design, the training recipe, the eval harness, the data boundary, the cooling system, and the procurement call. Today's stories mostly sit in those earlier layers.

56. Damra Vol 00:20:50

    Which is why the day feels operational rather than flashy. The model can be impressive and still depend on approval paths, cables, substations, and private-field handling. Those aren't side issues once the model is used by banks, governments, and agents with tools.

57. Lenar Kess 00:21:08

    So the day starts with Axios naming the people around White House AI decisions, and it ends with a pretty concrete builder checklist. If access can change by government pressure, design for regional and customer-specific fallback. If capacity depends on water and power, treat location as part of the compute plan. If agents use private documents, remove secrets before the model gets a chance to repeat them.

58. Damra Vol 00:21:33

    And if your eval only checks the final answer, it is under-reading the system. The SafeClawBench result is a useful warning there: the state after the tool call can disagree with the text you thought you had approved. Once the agent can write, send, store, or mutate, the evidence is outside the chat bubble.

59. Lenar Kess 00:21:52

    That is the sentence I would keep from the research block. The evidence is outside the chat bubble. And from the policy block, I would keep the companion sentence: the permission is outside the contract. Those two ideas are enough for one Thursday.

60. Damra Vol 00:22:07

    The practical test now is whether companies start documenting those external dependencies with the same seriousness they document model latency. Which jurisdictions can use the model? Which tools can mutate state? Which secrets are replaced with handles? Which data center regions can flex under grid stress? Those answers should be visible before the incident.

61. Lenar Kess 00:22:29

    And the next model launch that earns attention will probably answer some of those questions directly, not just with benchmark charts. Who can say no, who pays for the grid connection, and where the private fields go before the agent sees them: that is where the design work moved today. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic