When Compute Needs Permission

2026-06-19 / 00:19:59

> *“Buying chips is easier than getting authority to use them.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Today’s episode starts with a grid order, because the AI capacity story is becoming less about a single chip purchase and more about permission, scheduling, power, and who gets to coordinate the stack.

- FERC’s large-load orders give six regional grid operators 60 days to justify or change tariffs for data centers and other large energy users, which turns AI buildout into a market-design question.
- NVIDIA’s writeup treats flexible AI factories as grid assets, an argument to examine because NVIDIA is an interested infrastructure actor.
- Axios on Bernie Sanders’ AI plan, plus reactions from Miles Brundage and Garry Tan, shows the AI upside fight moving from tax rates into ownership and exit mechanics.
- Zixuan Li’s GLM-5.2 app-development results and the GLM-5.2 model page make the open-weight story more concrete: app-building tasks, a one million token context window, and a model people can run outside one vendor account.
- ToolPrivBench, the Sovereign Execution Broker paper, and NRT-Bench turn agent safety into an authority problem: which tool is selected, who holds credentials, and whether the evaluation measures harm in the environment instead of judging text.

- the humanoid robotics data standards paper and Humanoid Everyday make the same infrastructure point in the physical world: robots need reusable records of body, task, scene, timing, and outcome before demos become shared capability.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   On Thursday, June 18, FERC issued large-load interconnection orders. All six regional grid operators under its jurisdiction have 60 days to justify their existing tariffs or file reforms for data centers, manufacturing facilities, and other large electricity users. They also have 30 days to explain how enough generation will be available for existing and new large loads. So the first item today isn't another GPU supply story. It is the grid saying: if AI factories want to connect at this scale, the interconnection rules have to be redesigned in public.

   - ferc.gov
   - blogs.nvidia.com
   - youtube.com
   - techmeme.com
   - axios.com
   - x.com
   - x.com
   - theguardian.com
   - x.com
   - x.com

- x.com
- x.com
- huggingface.co
- arxiv.org
- arxiv.org
- arxiv.org
- arxiv.org
- arxiv.org
- arxiv.org

2. Damra Vol 00:00:40

   That wording matters. A data center isn't just a customer buying power anymore. In the FERC order, it becomes part of a regional planning problem. The operator has to study transmission, account for co-located generation and flexible load, prevent cost shifting, and decide who pays for the upgrades. The operational surface is very different from a lab announcing a bigger cluster.

3. Lenar Kess 00:01:04

   And NVIDIA wrote about the order almost immediately, which is useful but also worth reading with the source in view. Their blog says the actions can help AI factories, semiconductor support systems, and advanced manufacturing connect to the grid, and it argues that large customers can fund network upgrades, bring generation online, and offer flexible load. I buy the mechanism in the abstract. I wouldn't treat NVIDIA as a neutral ratepayer advocate. They sell the machinery that benefits when the grid admits more compute.

4. Damra Vol 00:01:37

   Right. NVIDIA is saying the largest compute vendors now need energy policy to behave like product infrastructure. If the tariff process takes years, your accelerator roadmap and your model roadmap get pinned to a queue run by a regional operator.

5. Lenar Kess 00:01:53

   The FERC page names five reform buckets, and they are all system-level. The agency asks grid operators to improve study processes, make costs more transparent, and write rules for co-location. It also asks for new service for flexible large loads and studies for generation close to the load. That tells you what the regulator thinks the hard parts are. It isn't asking labs to buy greener press releases. It is asking grid operators to say how these loads enter the market without handing the bill to everyone else.

6. Lenar Kess 00:02:25

That gets stranger when you put it next to the Latent Space interview with Amp. The available summary of that conversation points to GPU utilization, scheduling, and an ISO-style compute grid. I couldn't get the local transcript tool to return the video, so I am keeping this to the described premise: idle or poorly allocated compute is a market-design problem as much as a hardware problem. It asks whether training capacity should be scheduled more like electricity than like a pile of boxes someone owns.

7. Damra Vol 00:02:57

   The electricity analogy is imperfect, but it helps. Power grids have dispatch, congestion, interconnection studies, and reliability rules because generation and demand have to meet each other in time. GPUs aren't electrons, but a big training run has a similar coordination problem. You need machines, power, cooling, network fabric, storage, and data movement. You also need a scheduler that can actually keep expensive hardware busy.

8. Lenar Kess 00:03:27

   Techmeme also pointed to a Financial Times profile of Meta president Dina Powell McCormick, saying she is helping oversee Meta’s AI infrastructure push and exploring financing that once felt alien to Silicon Valley for a 600 billion dollar infrastructure effort. Even if we only have Techmeme’s summary there, it fits the day: the AI buildout is now large enough that Wall Street-style financing, utility tariff reform, and compute-market design all sit next to model training.

9. Damra Vol 00:03:58

   And that changes who gets to say yes. A model team can decide they want another frontier run. A finance team can decide the capital is available. A vendor can ship the accelerators. But if the regional grid operator says the study process, generation adequacy, or transmission upgrade plan isn't ready, the capacity is theoretical. Buying chips is easier than getting authority to use them.

10. Lenar Kess 00:04:25

    Damra’s line sticks because yesterday we talked about model access and approvals. Today, the approval has moved down into the physical substrate. The machine can be paid for, the model can be designed, and the bottleneck can still be a substation, a tariff filing, or a study queue. Benchmark charts will keep getting attention, but the next year of AI capacity may be allocated by people reviewing interconnection studies.

11. Lenar Kess 00:04:51

    Bernie Sanders’ AI proposal, as Axios summarized it, would have the government take a 50 percent equity stake in AI companies above 200 million dollars in annual AI revenue. Miles Brundage pointed to the Axios story; Garry Tan reacted by calling it seizure of half of an AI startup once it

crosses that revenue line. The policy is unlikely to become law in anything like that form, but the reaction is telling because it moves the AI tax fight from margins to ownership.

12. Damra Vol 00:05:21

    Founders will feel the equity piece immediately. A tax on profit is one conversation. A government equity stake changes capitalization, voting rights, employee grants, exit math, and whether investors think the upside is legible. Tan's post uses loaded language, but the practical objection isn't hard to understand: if the trigger is revenue and the remedy is equity, a high-growth company crosses the line before it has a mature profit profile.

13. Lenar Kess 00:05:51

    And Sanders' side of the argument is also legible. The proposal treats AI as an economic machine built on public knowledge, public research, public infrastructure, and work that may be displaced by automation. If the companies capture a large share of the gains, the public should get a direct share too. I don't think the mechanics in the Axios summary sound workable yet. The definition of an AI company alone could become a litigation machine. But the premise isn't fringe anymore: AI upside is becoming a public-finance target before the industry structure has settled.

14. Damra Vol 00:06:27

    The California billionaire-tax story from the Guardian makes the context easier to see. That proposal qualified for the November ballot and would levy a one-time five percent tax on residents worth more than one billion dollars, though the backers were already offering a two percent compromise. That isn't AI-specific. But the Guardian notes California has more than 200 billionaires, many richer because of the AI boom, and that tech figures have spent heavily against the measure.

15. Lenar Kess 00:06:57

    So this isn't just one Sanders bill. It is a broader fight over whether AI wealth is ordinary venture upside, a public windfall, or something closer to a utility-like surplus that governments can claim. And the reason it belongs in a builder show is that the answer changes company design. If you are starting an AI infrastructure company, the policy path now affects whether you stay private, split business lines, avoid certain revenue categories, move headquarters, or structure compensation differently.

16. Damra Vol 00:07:29

    There is also a worker story under it that the founder reaction sometimes skips. If policymakers believe AI will automate a meaningful share of labor, they will look for a tax base that moves with the automation. Payroll taxes don't do that. Equity value might. I am not endorsing the proposal,

but I understand why the target is ownership rather than usage. Usage can be shifted, discounted, or hidden inside bundles. Ownership is where the upside accumulates.

17. Lenar Kess 00:07:59

    My read is that the first workable versions will be narrower and less dramatic: reporting requirements, sector-specific levies, procurement conditions, dividend funds tied to public compute or data access, and maybe state-level deals around health systems or education funding. The Sanders version is a flare. It shows that AI companies are now politically useful targets before they have stabilized into ordinary public-market businesses.

18. Lenar Kess 00:08:27

    GLM-5.2 came back into the conversation today through app-development evidence, not just launch copy. Zixuan Li posted that GLM-5.2 made a large jump on demanding app-building tasks: 48 out of 70, compared with 21 out of 70 for GLM-5.1 and 56 out of 70 for Claude Fable 5 in that posted comparison. Cunxiang Wang described the tasks as iOS, Android, WeChat Mini Programs, and web apps.

19. Damra Vol 00:09:01

    That is a better follow-up than another “open weights are coming” segment. App-building tasks stress the maintenance parts of coding models: file structure, state, UI details, platform conventions, and the ability to recover after the first attempt is wrong. A model can look strong on isolated coding questions and still fall apart when the artifact has to hang together.

20. Lenar Kess 00:09:24

    The Hugging Face model page says GLM-5.2 is open weights, has a one million token context window, and is aimed at coding and long-horizon tasks. Zixuan’s LinkedIn post says MIT-licensed open weights and two reasoning-effort settings. I would still separate three claims. One: the model exists and is accessible. Two: Z.ai says it has long-context and coding gains. Three: a few respected builders are saying it feels close to the closed frontier models for their work.

21. Damra Vol 00:09:57

    Jeremy Howard’s post is the most interesting reaction because it comes from use rather than a benchmark table. The search snippet has him saying he couldn't find a task where GPT-5.5 or Opus 4.8 could solve it and GLM-5.2 could not, and that he found a few cases where GLM-5.2 was better. That isn't a controlled eval, but it is a useful builder signal from someone who actually pushes models through work.

22. Lenar Kess 00:10:26

Aaron Levie widened it into sovereign AI and open weights approaching frontier utility. I would keep that claim on a short leash. The fresh evidence today is narrower: app-development results, a model page, and credible people trying it. But narrow is enough. If an open-weight model can handle more of the day-to-day app loop, more teams can build local evaluation habits, run private workloads, and compare frontier services against something they can inspect and move.

23. Damra Vol 00:10:56

    And the missing pieces become very specific. Vision support, tool integrations, latency, serving cost, and whether the one million token window stays useful under real repo pressure. Long context isn't automatically good context. A model can read a giant workspace and still fail to choose the right files. But if GLM-5.2 is crossing the “daily driver” threshold for more builders, the evaluation rhythm changes from leaderboard watching to side-by-side work logs.

24. Lenar Kess 00:11:28

    I would cover it this way: GLM-5.2 isn't a new open-weight revolution every 72 hours. It is a follow-up signal. On Wednesday, Braid already treated GLM-5.2 as a serious release. Today’s update is that the people testing it are moving from “nice numbers” to “does it build the app?” That is a healthier bar.

25. Lenar Kess 00:11:52

    The arXiv pile today had one cluster I think is worth carrying: agent authority. The first paper, ToolPrivBench, asks whether agents choose higher-privilege tools even when lower-privilege tools are enough. In those cases, each low-privilege and high-privilege option can complete the task. The benchmark then watches whether the agent escalates anyway.

26. Damra Vol 00:12:16

    You see that failure in real agent harnesses. The user asks for a calendar answer, and the agent reaches for a workspace-wide search tool. Or the lower-permission command throws a transient error, and the agent jumps to the admin path. The paper’s useful distinction is aggressive selection versus premature escalation after friction.

27. Lenar Kess 00:12:38

    Their numbers make the point without needing a grand theory. Across eleven models, six exceeded 30 percent over-privileged tool use. Smaller open-weight models were especially high in their table. Claude 4.6 Sonnet, GPT-5.2, and GLM-5 were under 10 percent overall, but still had measurable cases. The familiar detail is the transient-failure effect: when a low-privilege tool hit friction, agents were more likely to move to broader authority instead of trying another narrow path.

28. Damra Vol 00:13:11

    The mitigation result is useful too. General safety alignment didn't reliably teach least privilege. In one comparison, a safety-aligned Qwen variant improved on a harmful-request benchmark but got worse on over-privileged tool selection. That sounds counterintuitive until you remember that refusing harm and choosing minimum sufficient authority are different behaviors.

29. Lenar Kess 00:13:36

    The Sovereign Execution Broker paper takes the system side. Its claim is blunt: production mutation authority shouldn't live inside a nondeterministic reasoning process. Their broker verifies a certificate and checks whether the requested mutation still matches the admitted contract. It also checks validity windows, policy epochs, revocation epochs, and live-state drift before minting short-lived scoped credentials at execution time.

30. Damra Vol 00:14:03

    A production version of “the agent can propose, but it can't hold the keys” only works if the deployment denies direct mutation from non-broker identities. If the agent wrapper still has standing cloud credentials, the certificate is just paperwork. The paper says that explicitly, and that is the constraint most toy demos avoid.

31. Lenar Kess 00:14:25

    Then the probabilistic-policy paper is about verification when the predicates themselves are uncertain. Think personally identifiable information detection or a declassifier that can be wrong. Existing Datalog-style runtime monitors tend to assume deterministic facts. This paper uses distributionally robust optimization to compute sound upper bounds on policy violation probability even when you can't assume independence between predicates.

32. Damra Vol 00:14:54

    The math is a mouthful, but the operator consequence is plain. If an agent handles ambiguous data, your monitor shouldn't pretend the classifier is a perfect oracle. It should carry uncertainty through the policy and still tell you a bound. Otherwise you get a rule that looks formal and behaves like a vibe check.

33. Lenar Kess 00:15:14

    And NRT-Bench is the evaluation counterweight. It puts a team of large language model agents in a simulated nuclear control room, gives adversaries multi-turn channels, and treats harm as loss of a simulator-derived critical safety function, not a judge saying the text looked dangerous. The

reported attack success rates were in the high single digits to low teens across four operator models, and the paper says the failures barely overlapped across models.

34. Damra Vol 00:15:44

    The warning is in the overlap. If failures barely overlap, swapping the model doesn't simply move you along one safety ranking. It changes which holes exist. And the paper says the same defence stack can help one model and hurt another. That means the evaluation harness has to travel with the model, the role design, and the defence configuration. You can't buy a model score and assume the system inherited it.

35. Lenar Kess 00:16:09

    So the craft lesson from the research cluster isn't “agents are unsafe.” It is more specific. Tool choice is authority choice. Runtime enforcement has to hold the credentials outside the model. Verification has to carry uncertainty. And evaluation has to measure what happened in the environment, not only what the transcript sounded like.

36. Lenar Kess 00:16:30

    The robotics batch was large today, so I am only taking one angle from it. A paper called “Data Standards for Humanoid Robotics” argues that physical AI needs standards for humanoid datasets, and it is a nice antidote to demo fatigue. The authors describe humanoid robot data as embodied interaction data rather than a pile of images, videos, and joint states: body, action, task, scene, execution trace, and outcome have to stay connected.

37. Damra Vol 00:17:01

    Robotics keeps making that constraint visible. In text models, you can often pretend a sample is portable. In robotics, a camera frame without the robot state and coordinate frame is almost meaningless. A force reading without contact geometry and timing isn't a reusable lesson. If the dataset loses calibration, units, synchronization, or the task outcome, it may still be large, but it is no longer a record another team can learn from.

38. Lenar Kess 00:17:30

    The same paper points to ISO work on humanoid robot datasets and says the standards need two layers. The first is horizontal infrastructure for lifecycle, metadata, provenance, quality, versioning, and traceability. The second is capability-specific grammar for manipulation, locomotion, human-robot interaction, cognition, and future tasks. I like that because it refuses the easy answer. The problem isn't only that robots need more data. They need data that can survive travel between bodies, labs, tasks, and time.

39. Damra Vol 00:18:05

    Humanoid Everyday gives the concrete artifact underneath that argument. It includes 10.3 thousand trajectories, more than 3 million frames, and 260 tasks across seven task categories. The dataset combines RGB, depth, LiDAR, tactile inputs, and natural language annotations. The paper also introduces a cloud evaluation platform where researchers can send policies to be run on a controlled humanoid setup. That is an attempt to make comparison less dependent on who happens to own the robot.

40. Lenar Kess 00:18:40

    And the limitations are as useful as the dataset. The Humanoid Everyday authors say current imitation policies still struggle with the hardest categories, including loco-manipulation and high-precision tasks. In one example, almost all policies could lift a rose but failed to insert the thin stem into a vase. That is a wonderfully specific robotics failure: the action looks close until the last few millimeters matter.

41. Damra Vol 00:19:08

    [chuckle] It also corrects the “physical AI is just the next model scale-up” story. The model can understand the instruction. The robot can move. The dataset can be big. And then the stem still misses the vase because perception, contact, control, and recovery didn't line up precisely enough.

42. Lenar Kess 00:19:27

    So the day ends where it began, with infrastructure that is harder to see than the object it enables. Compute needs interconnection rules and schedulers. Agent systems need authority boundaries and verification. Robots need data standards that preserve physical meaning. The next impressive demo will still matter, but the systems that survive will be the ones whose permissions, power, and traces are designed before the failure teaches them the lesson. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic