The Lease Needed Power

2026-07-06 / 00:27:18

> *"The AI buildout is being measured in leases, rack dates, grid queues, and rules about what synthetic characters are allowed to do."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Today's episode follows the AI buildout as it leaves the slide deck and runs into financing, rack design, grid queues, and product rules. The largest story is a set of infrastructure receipts: a Treasury warning, a reported Nvidia rack delay, Anthropic's Kentucky lease, and a Scottish data-center project whose renewable-power promise does not appear to survive contact with local evidence.

- Techmeme's NOTUS item on a draft Treasury report puts market-risk language next to the AI capex boom, with dotcom-era comparisons entering the official vocabulary.
- Techmeme's CNBC item on Nvidia Kyber NVL144 tracks a reported 12-month-plus delay and cancellation of NVL72x2, which makes rack-scale interconnects part of the AI timeline.
- Techmeme's CNBC item on Anthropic and TeraWulf gives the other side of the same story: a 20-year, $19 billion Kentucky data-center lease with about 400 megawatts of planned capacity.

- The Guardian's Lanarkshire investigation shows how quickly a national AI growth-zone promise turns into land, planning, grid access, and renewable-power math.
- Techmeme's China agent-rule item and Rest of World's web-novel reporting show Chinese platforms narrowing what users can create with AI before regulators and readers force the issue harder.
- Armin Ronacher's Claude tool-call critique, Fly.io's agent sandbox post, and Harrison Chase's harness note turn the builder segment toward model-harness contracts rather than demo polish.
- Forbes on Sam Altman's global-referee proposal, Forbes on FTC bias disclosure, and The Guardian on the FCA Mills review give three different versions of AI governance: standards, disclosure, and financial-services supervision.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Techmeme's top AI infrastructure items this morning put four different facts next to each other: a draft Treasury report warning about AI-market risk, a reported Nvidia rack delay, Anthropic signing a huge Kentucky data-center lease, and a Guardian investigation into a Scottish AI project whose

renewable-power promise doesn't appear to hold up. [pause] Those aren't four versions of one story. They are four places where the AI buildout meets a different test. A finance warning becomes credit discipline. A rack design meets the factory. A lease needs megawatts on a schedule. And a political promise about green data centers runs into land, grid access, planning approval, and cables in the ground.

- techmeme.com
- techmeme.com
- techmeme.com
- theguardian.com
- techmeme.com
- restofworld.org
- techmeme.com
- x.com
- fly.io
- forbes.com
- forbes.com
- forbes.com
- theguardian.com
- theguardian.com
- techmeme.com
- techmeme.com
- theguardian.com
- chinatalk.media
- lucumr.pocoo.org

2. Damra Vol 00:00:50

   The order matters there. The fantasy version of the AI boom starts with models and then assumes everything around them catches up. Today's source list starts with the physical world asking for receipts. Can the rack ship? Can the site draw power? Can the lease turn into usable capacity before the customer's model roadmap needs it? Can the money wait that long without getting nervous?

3. Lenar Kess 00:01:14

   The Treasury item is via Eric Katz at NOTUS, surfaced on Techmeme. The NOTUS report says a draft US Treasury Department document is set to warn about AI-market risks and compare parts of the current moment to the dotcom crash. That's not a crash call. I wouldn't treat it that way. It is a sign that the official financial vocabulary is catching up to the scale of the spend. Once the government starts writing about AI capex with dotcom-era comparisons, the conversation has

moved from ‘can these companies raise enough money?’ to ‘what happens if a lot of these promises are capitalized before the returns arrive?’

4. Damra Vol 00:01:51

   That creates a different kind of pressure than a product delay. A product delay hurts the roadmap. A financial-risk memo changes who has permission to be skeptical. You can have a CFO, a regulator, or a lender say, ‘show me the signed customers, show me the power contract, show me the delivery date,’ and that request no longer sounds like they don’t understand AI. It sounds like they read the same memo.

5. Lenar Kess 00:02:17

   Then you have the Nvidia item. CNBC, citing SemiAnalysis, reports that Nvidia’s Kyber NVL144 rack-scale system has slipped by more than 12 months to 2028 because of printed-circuit-board manufacturing issues, and that the NVL72x2 architecture has been canceled. The component that moved the date matters more than the date itself. This isn't a model waiting for another benchmark pass. It is the physical rack system around Rubin Ultra. Scale-up interconnects have to work, copper and co-packaged optics have to meet the design, and the midplane has to be manufacturable before that promised compute tier exists.

6. Damra Vol 00:02:59

   The point that stuck with me is that a printed circuit board becomes a model-timeline event. If Kyber is delayed because the midplane is hard to manufacture, then the AI roadmap is no longer only about chip supply. It’s about whether the rack-scale architecture can move enough data across enough accelerators without becoming a science project. The model people can be completely right about demand. They can still end up waiting on electrical engineering paperwork until the schedule breaks.

7. Lenar Kess 00:03:27

   Right. And in the same morning window, Techmeme has CNBC reporting that Anthropic signed a 20-year, $19 billion lease with TeraWulf for a Kentucky data center expected to reach roughly 400 megawatts of capacity, with first power delivery in the second half of 2027. That isn't a vague cloud partnership. It is a long lease on a power-hungry physical site. The supplier is named, the state is named, the capacity figure is named, and the first-power window is part of the report.

8. Damra Vol 00:03:59

   It also tells you why the Treasury-style skepticism and the Nvidia-style schedule risk belong in the same conversation, even if they aren't the same claim. Anthropic can have extraordinary demand. It

can also need a site that delivers hundreds of megawatts on a timeline that lines up with its product and training plans. The lease is evidence of seriousness, not evidence that the power already exists.

9. Lenar Kess 00:04:23

   That distinction helps with the Guardian's Lanarkshire reporting, which is exactly about the gap between a promise and a site. Aisha Down reports that the UK government announced an 8.2 billion pound AI data-center complex in Lanarkshire in January, with CoreWeave and DataVita attached, and said it would be powered entirely from on-site renewables and built by 2030. The Guardian says documents obtained through freedom-of-information requests and public-record analysis show the site has no prospect of meeting that goal.

10. Damra Vol 00:04:54

    And the reported private acknowledgement is blunt. The government and developers were publicly talking about up to one gigawatt of new energy infrastructure while privately acknowledging an issue with power provision. That doesn't mean the project can't exist. It means the public story of the project depended on a power story that the paperwork didn't support.

11. Lenar Kess 00:05:16

    The Guardian's numbers make the promise feel less abstract. DataVita's plan includes 400 megawatts of solar and 800 megawatts of wind. Guardian analysis says the renewable buildout would need somewhere between 40 and 100 square kilometers of land. The planning applications currently on file cover about 2 square kilometers, and DataVita's own site claims over 1,000 acres, or about 4 square kilometers, of renewables. So even before you get to grid queues, the physical footprint edits the press release.

12. Damra Vol 00:05:48

    That's a good place to keep the altitude low. This isn't the AI bubble popping in one Scottish village. It is a reminder that national AI strategy now contains a lot of civil-engineering claims. If those claims are wrong, the error doesn't stay inside the AI sector. It competes with housing, hospitals, industrial power, and every other project sitting in the same queue.

13. Lenar Kess 00:06:11

    Exactly. I also don't want to flatten the good news out of this. The Anthropic lease is demand meeting supply planning. The Nvidia delay, if the report holds, is a technical constraint rather than a demand collapse. The Treasury warning is a warning, not a verdict. The Guardian investigation is about one project's claims, not every data center in Europe. All four items make the same habit harder: treating AI infrastructure as a future state that can be described before it is built.

14. Damra Vol 00:06:43

    And that habit has been useful for raising money. [pause] It may even be useful for getting political attention. But the next proof points are physical and specific. Power has to be delivered first. Racks have to be available. Capacity has to be signed. Grid-connection dates and planning documents have to match the public claim. The companies that can show those receipts will sound different from the companies that only have renderings and round numbers.

15. Lenar Kess 00:07:12

    There's a callback to Saturday's Braid episode here, but I want to keep it narrow. We talked then about data-center constraints and site visits. Today's update is that the same constraint is showing up in four registers at once: financial supervision, rack manufacturing, leased capacity, and local power evidence. That doesn't make the story more dramatic. It makes it harder to hand-wave. The buildout is still happening. It is also getting measured by people who aren't grading on model-demo curves.

16. Damra Vol 00:07:42

    The phrase I'd use is measurement friction. The model demo can travel instantly. The data center has to ask permission from physics. It also has to satisfy neighbors, lenders, and the grid operator. That doesn't kill the project. It changes which claims survive the calendar.

17. Lenar Kess 00:08:00

    Techmeme's China item says ByteDance's Doubao and Alibaba's Qwen will disable humanlike and user-created agents before China's anthropomorphic AI interaction rules take effect on July 15. That is a concrete product change. Users don't just lose access to a model capability in the abstract; they lose a particular way of making synthetic characters or agents behave like people.

18. Damra Vol 00:08:25

    And that makes the regulation feel less like a model-access rule and more like an interface rule. The state isn't only asking who can run a powerful model. It is asking what kind of relationship the product is allowed to simulate. That is a much more intimate boundary. It reaches into companions, role-play, customer-service characters, and user-built agents that might otherwise look harmless because they sit on top of a general model.

19. Lenar Kess 00:08:51

    Rest of World has the matching culture-side story. Viola Zhou reports that Chinese web-novel platforms owned by Tencent, ByteDance, and Baidu first embraced AI writing tools. Now they are clamping down after reader backlash, plagiarism worries, and a flood of low-quality automated

fiction. The numbers are useful. A civil engineer named Gordon Sheng used DeepSeek and an AI writing tool to generate a short story in five minutes, and Rest of World says it drew more than 5,500 reads in 10 days on Tomato Novel.

20. Damra Vol 00:09:25

    That’s the seduction. Five minutes to a story with readers. For someone who has ideas and no writing habit, that isn't trivial. But the platform problem appears as soon as everyone can do it. The reader is no longer just choosing among stories. The reader is trying to detect whether the story has been bulk-produced, whether it accidentally left a prompt in the middle of a chapter, and whether the writer borrowed somebody else’s plot with a layer of synthetic rewriting on top.

21. Lenar Kess 00:09:54

    The platform responses are specific. Rest of World says Jinjiang’s founder asked authors to use AI only for research and proofreading and asked readers to report suspected AI writing. Tomato Novel capped how many words each account can publish per day. In June, Tomato Novel rejected more than 104,000 low-quality submissions, including AI-written ones, according to the company statement Rest of World cites.

22. Damra Vol 00:10:20

    That word cap is such a platform-native solution. It doesn't decide whether a paragraph has a soul. It changes the economics of flooding the market. And it shows the same basic pressure as the agent rules: once synthetic generation becomes an interface for ordinary users, the platform has to govern volume, persona, and imitation, not just model safety in a lab.

23. Lenar Kess 00:10:45

    There’s a temptation to make this a broad China-versus-US story, but the sources don't need that. The narrower point is stronger. China-facing AI products are being constrained from several directions at once: rules about anthropomorphic interaction, platform rules about AI fiction, reader backlash against low-quality generation, and author anxiety about training rights and plagiarism. These are product surfaces. They are where policy, culture, and user behavior meet.

24. Damra Vol 00:11:16

    And the creative opening is still there. Sheng’s quote in Rest of World isn't cynical. He says many people have the spark for a story and don't know how to express it. AI closed that gap for him. The conflict starts when that benefit is multiplied through a platform whose business depends on attention, adaptation rights, and reader trust.

25. Lenar Kess 00:11:36

I like pairing these two items because one is about AI agents that feel human, and the other is about AI prose that feels cheap to readers when it arrives in bulk. Both cases force platforms to decide which humanlike affordances they want to encourage, which ones they want to throttle, and which ones regulators will no longer let them offer.

26. Damra Vol 00:11:56

    The July 15 deadline gives the agent story a date. The web-novel backlash gives it a texture. Users don't wait for a policy memo to decide a platform feels worse. Sometimes they just start taking screenshots of the prompt that got left inside the chapter.

27. Lenar Kess 00:12:12

    Armin Ronacher published a post called ‘Better Models: Worse Tools’ after debugging a strange failure in Pi’s edit tool. Newer Claude models, including Opus 4.8 and Sonnet 5, sometimes called the edit tool with correct old text and new text, but then added invented fields inside the nested edits array. Armin lists invented names for uniqueness, matching, forced match counts, and second old/new text fields. Pi rejected those calls because the schema didn't allow them.

28. Damra Vol 00:12:45

    That is a wonderfully annoying bug because the model understood the task. It found the right text. It wrote the right replacement. Then, at the end of the object, it hallucinated a little handle on the tool call and broke the contract. The failure isn't intelligence. It is sampling pressure at the boundary between a model’s learned tool habits and somebody else’s stricter schema.

29. Lenar Kess 00:13:08

    Armin’s reproduction detail is important. A fresh single-turn prompt didn't trigger it. It showed up after an agentic history where the model had read files, diagnosed a problem, and composed a multi-line edit. In one user’s transcript, continuing the session made Opus 4.8 fail around 20 percent of the time. Stripping thinking blocks from history cut the rate by half. Turning on strict tool invocation eliminated it in his runs.

30. Damra Vol 00:13:37

    That puts pressure on a comforting idea a lot of us carry around: that a tool schema is a neutral contract and the model will just follow it if the instructions are good. Armin’s read is that newer Anthropic models may have been post-trained in a Claude Code-like environment, where the edit tool is flatter and the client repairs a lot of malformed input. If a forgiving harness absorbs the mistake, the model can get rewarded for completing the task even when the raw call is sloppy.

31. Lenar Kess 00:14:06

He points at Claude Code's client as evidence that the harness is forgiving. It accepts parameter aliases and type coercions. It repairs Unicode, retries some failures, filters unknown keys, and has no strict mode. I'm keeping the inference bounded because Anthropic's training environment isn't public, but the observed bug is concrete. A model can improve at the provider's own tool ecology and become worse for a tool with a different shape.

32. Damra Vol 00:14:34

    That is the builder story for me. The harness is no longer a wrapper you can treat as interchangeable. It is part of the distribution the model learned. If the dominant harness tolerates aliases and filters extra keys, everyone else building a stricter harness has to decide whether to become more like the dominant environment or force the model into stricter decoding. Either way, the harness is in the product now.

33. Lenar Kess 00:15:01

    Harrison Chase had a short X note about 'agent harnesses' versus 'agent frameworks,' and that wording fits this bug. A framework sounds like a thing you import. A harness sounds like the runtime around the model: message history, tools, retries, state, validation, observation, memory, approvals, and the way errors get fed back. Armin's bug lives inside that runtime, not inside a single API call.

34. Damra Vol 00:15:28

    The Fly.io post gives the same idea a more physical feel. Daniel Botha writes about agents that run risky commands inside Sprites, with a durable agent process separate from the disposable place where commands execute. The examples are very concrete: an agent writes two migration files, a checkpoint is taken, then a cleanup prompt causes the model to delete the app and its toolchain. In a Sprite, Fly says the checkpoint restore brings the files and git back in about nine seconds.

35. Lenar Kess 00:15:59

    And the design choice there isn't 'tell the agent to be safer.' It is 'make the agent's home and the agent's execution environment different places.' The agent can keep memory, history, and skills in one durable loop, while commands run somewhere disposable or resumable by task. That pairs nicely with Armin's strict-tool lesson. Some reliability comes from better model behavior. Some reliability comes from refusing to make model behavior carry the whole safety story.

36. Damra Vol 00:16:29

    The local-model item from the LocalLLaMA subreddit points in the same direction, even if I would keep it as background. The LocalLLaMA post says llama-server can throw away useful key-value cache state under certain long-context conditions. That is a narrower bug, but it rhymes technically:

the model's apparent capability depends on what the runtime preserves, reuses, validates, and throws away. If the cache is wrong, the long context isn't the long context you thought you bought.

37. Lenar Kess 00:17:01

    Forbes also has a category piece arguing that agent gateways are becoming a control layer for enterprise AI. I don't want to overstate that one because it is more analysis than a primary release. It belongs in the segment because enterprises are going to ask for the same things Armin and Fly are circling from different ends. They need policy and audit. They need permissions and routing. They need retry behavior, tool access, and a place to enforce contracts between model calls and real systems.

38. Damra Vol 00:17:30

    The funny part is that agent tooling used to sell itself as freedom from workflow software. Just give the model tools and let it work. Now the grown-up version is full of validators, checkpoints, state stores, sandboxes, gateways, and constrained decoding. I don't mean that as a complaint. Demos grow into workflows only when the runtime can absorb mistakes that the model will keep making.

39. Lenar Kess 00:17:55

    My read is that the next serious agent comparison will spend less time asking which model is smartest. It will ask what tool-call shapes the model follows under pressure, which harness the model appears to know, how it behaves after a long transcript, what strict mode costs, and which runtime pieces can be inspected when something breaks. That is less glamorous than a benchmark table, but it is closer to where the strange failures happen.

40. Damra Vol 00:18:21

    And it keeps the optimism intact. A harness is a place to put craft. If tool calls are text, and schemas have distributional distance, and checkpoints make risky execution reversible, then builders have levers. They just aren't all inside the model card.

41. Lenar Kess 00:18:37

    Forbes has Anisha Sircar reporting that Sam Altman wants a US-led global standards body for AI and, in the same broad governance conversation, is open to the US government owning a piece of OpenAI. That combination is striking because it puts referee language and ownership language very close together. One is about setting rules. The other is about the state having a direct stake in one of the companies being ruled.

42. Damra Vol 00:19:02

That is a hard pairing to make boring. [pause] A global standards body can sound like neutral infrastructure for safety and interoperability. Government equity sounds like industrial policy, bailout logic, or strategic alignment, depending on who is describing it. Put them together and the governance proposal starts to look less like a referee standing outside the game and more like a government deciding which parts of the game count as national capacity.

43. Lenar Kess 00:19:30

    Today's governance items aren't coordinated. They are crowded. Forbes has a Lance Eliot piece on an FTC policy idea aimed at AI makers disclosing the truth about bias in large language models. The Guardian has Kalyeena Makortoff on the FCA's Mills review, which says AI will reshape financial services by 2030 and could improve access and personalization while amplifying fraud, cybersecurity, consumer-harm, and market-concentration risks.

44. Damra Vol 00:20:00

    The FCA piece is the most operational of the set. The review recommends that the regulator use its own AI-enabled model to supervise firms, and that the government expand the FCA's authority over critical third parties such as AI firms and cloud providers. That isn't a grand theory of AI safety. It is financial supervision saying, 'if the bank's consumer interface depends on a model or a cloud provider, our current perimeter may be too small.'

45. Lenar Kess 00:20:28

    And then the Guardian opinion piece by Bruce Schneier and Jon Penney is a different category again. It isn't a government action; it is an argument about AI-powered surveillance policy. I'd include it only lightly because it helps show the governance surface expanding. One lane is global standards. Another is model-bias disclosure. The FCA lane is fraud and third-party concentration in finance. The Schneier and Penney lane is civil-liberties risk from surveillance systems. Those don't resolve into one neat rulebook.

46. Damra Vol 00:20:59

    They also create different winners. A standards body helps whoever can afford to sit at the table and implement the standard. Bias disclosure helps firms that already have evaluation and documentation machinery. Financial-services oversight pulls cloud and AI providers into a regulated supply chain. Surveillance limits, if they arrive, would constrain government buyers and vendors in a different way. AI governance is becoming plural because AI is entering plural institutions.

47. Lenar Kess 00:21:32

That pluralism is why I’m wary of the single-referee story, even when I understand the appeal. The systems are too varied now. A chatbot companion, a bank-advice agent, a police surveillance tool, a coding model, and a data-center finance vehicle don't need the same referee. They need rules that match the harm, the dependency, the incentive, and the institution using the system.

48. Damra Vol 00:21:56

    And yet companies will keep asking for harmonization because fragmented rules are expensive. That doesn't make them wrong. It means standards become a power question. Who gets to define the test, who pays for compliance, who is too small to keep up, and who can turn a compliance department into a competitive advantage?

49. Lenar Kess 00:22:15

    The Altman item is useful because it forces that question into the open. A US-led referee can be a way to prevent a race to the bottom. It can also be a way to write the rules around the firms with the most capital, the deepest government relationships, and the largest infrastructure commitments. I don’t think those are mutually exclusive. That is the whole tension.

50. Damra Vol 00:22:37

    And the FCA story brings it back to the person at the other end of the system: someone getting financial advice, credit, fraud detection, account support, or insurance triage through an AI-enabled service. For them, the standards-body argument is distant. The question in their life is whether the regulator can see the system clearly enough to intervene when the model, the vendor, or the bank’s automation hurts them.

51. Lenar Kess 00:23:04

    Two shorter items before we close. First, Techmeme’s security lead points to BleepingComputer and Sysdig reporting JadePuffer, described as the first known agentic ransomware operation. The claim is specific: the operation exploited CVE-2025-3248 in an internet-facing Langflow instance, then used an AI agent to adapt in real time, retry steps, and run an end-to-end extortion campaign against databases.

52. Damra Vol 00:23:32

    That is a good example of a story where the noun can run ahead of the evidence. ‘Agentic ransomware’ sounds like a movie poster. The useful part is narrower: retrying and adapting inside a known intrusion path changes tempo. A human still appears to have selected or initiated the operation in the reports we saw, but the agent can carry more of the messy middle at machine speed. Incident response teams care about that middle.

53. Lenar Kess 00:23:59

    The technical root is familiar too. The cited CVE is in Langflow versions before 1.3, according to the National Vulnerability Database listing Techmeme links. So this isn't magic malware appearing from nowhere. It is an exposed, vulnerable AI-adjacent service, followed by automation that can steal credentials, expand access, and push toward extortion. The novelty is the handoff between old intrusion hygiene and newer agent behavior.

54. Damra Vol 00:24:27

    And that makes it more useful than a generic ‘criminals will use AI’ warning. You can inspect exposed Langflow instances. You can patch. You can watch for credential harvesting. You can ask whether your detection assumes a human pace between steps. The agent part is new; the doorway is painfully familiar.

55. Lenar Kess 00:24:46

    The other brief item is robotics. Techmeme has TechCrunch’s Q&A with Agility Robotics CEO Peggy Johnson about the company going public through a SPAC and treating the physical layer as its proprietary advantage. The Guardian has Amy Hawkins in Beijing on China’s push to solve robotic hands, and ChinaTalk has a Unitree-centered conversation about why Chinese robotics capacity matters. This is a supporting beat today, but the details are good.

56. Damra Vol 00:25:15

    The Guardian hand story is the one with texture. LinkerBot’s founder, Zhou Yong, says making a robotic hand is one hundred times more difficult than making a humanoid because the dexterity is ten times other body parts while the volume is one tenth. The company says it makes about 5,000 hands a month and wants to double that. That moves the robotics conversation away from dancing humanoids and toward manipulation. Can the machine button a shirt, pack groceries, handle pressure, feel touch, and learn from human motion data?

57. Lenar Kess 00:25:49

    ChinaTalk’s Unitree discussion pushes the market side: China’s advantage in vertical integration, actuator manufacturing, and supply chains that can make robots cheaper. The strongest line there, to me, is the warning that you can't AI your way out of a hardware problem or a supply chain that makes the robot too expensive. That sits nicely next to today’s data-center story, without needing to merge them. Physical AI has its own version of the same discipline: the model can be exciting, and the bill of materials can still decide the market.

58. Damra Vol 00:26:22

I like that we end on hands because it is almost comically concrete. After data-center leases, tool schemas, global referees, and ransomware agents, the future still needs fingers that can feel pressure without crushing an egg. That is a nice correction to the mind-only version of AI. Intelligence keeps asking for a body, a grid connection, a sandbox, a regulator, or a reader who still wants to read the next chapter.

59. Lenar Kess 00:26:51

    Tomorrow, I'd trust the items that attach claims to artifacts: a delivery date, a filing, a schema, a patch, a planning document, a sourced report, or a robot hand doing something more demanding than waving. Today's episode made the systems around the model easier to see. They leave paperwork, logs, invoices, rejected submissions, and sometimes 5,000 hands a month. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic