

US Orders DeepSeek Out of Federal Agencies as DeepSeek Open-Sources Its Inference Engine

2026-10-20 / 00:22:15

“Weights are a file. Getting tokens out of them cheaply is an engineering problem, and DeepSeek just published its answer.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

DHS has ordered DeepSeek out of federal civilian systems, including products built on DeepSeek implementations, on the same day DeepSeek published the serving stack it runs in production. Around that sit a \$35 billion chip order, Anthropic's leaked bill, a leaked Army agent program, and Goldman's argument that power is the bottleneck.

- [SiliconANGLE on the DHS order](#): the Federal Register language reaches third-party products built using DeepSeek implementations, which is where procurement offices get stuck.
- [DeepInfer on GitHub](#): DeepSeek's production inference stack, with throughput claims that so far come only from its own README.
- [Reuters via Techmeme on the RTX PRO 5500 order](#): more than three million export-compliant cards for ByteDance, Alibaba, and Tencent, per sources.
- [ChinaTalk, The Lithography Lever](#): the case for going upstream to DUV machines when card counts can't be capped.
- [Transformer via Techmeme on Anthropic's leaked bill](#) and [Axios on the self-rule experiment](#): a weaker-than-FRONTIER draft next to a White House line that incident

notification is not optional.

- The Intercept via Techmeme on ECHOTRIBBLE: a nascent Army Cyber Command program for agents that spread across networks.
- Axios on Goldman's power report: 76 of 86 cancelled or delayed projects trace to local opposition.
- Jedidiah Katz's AI Engineer talk: why cost per task diverges even when eval scores converge.

SEGMENTS

<u>00:00:04</u>	DHS orders DeepSeek out of federal systems
<u>00:06:13</u>	A \$35B export-compliant GPU order and other China-US pressure
<u>00:08:54</u>	Anthropic's leaked federal bill and the self-governance experiment
<u>00:11:19</u>	ECHOTRIBBLE and the containment side of agents
<u>00:13:18</u>	Goldman on power as the forever bottleneck
<u>00:16:03</u>	Do smarter models kill the harness?
<u>00:18:40</u>	The brief: deals, fast models, COBOL, robots, and Meta Muse

Transcript

1. Lenar Kess 00:00:04

Picture running IT at a federal agency. On Tuesday morning you open an order from the Department of Homeland Security saying DeepSeek has to come out of your systems. That's easy if somebody installed the DeepSeek app. But what if nobody did, and the document summarizer you bought from a vendor two years ago calls a DeepSeek model underneath?

- siliconangle.com
- techcrunch.com
- github.com

- techmeme.com
- aws.amazon.com
- techmeme.com
- chinatalk.media
- techmeme.com
- techmeme.com
- techmeme.com
- techmeme.com
- axios.com
- courtlistener.com
- techcrunch.com
- techmeme.com
- indianexpress.com
- indianexpress.com
- axios.com
- techmeme.com
- neowin.net
- youtube.com
- youtube.com
- techmeme.com
- inceptionlabs.ai
- openai.com
- cnbc.com
- tomshardware.com
- arstechnica.com

2. Damra Vol 00:00:21

That's the version I keep turning over, because going by the quoted language, the order reaches there. It isn't only DeepSeek's own products. So how would you even find out?

3. Lenar Kess 00:00:31

That's the lead today. We'll start with the ban and what its language covers, then the odd pairing of DeepSeek open-sourcing the software it uses to serve its models. After that come the China and US pressure points that all showed up within a day of each other, including a thirty-five billion dollar chip order. Then Anthropic's leaked federal bill and a leaked Army program for agents that spread across networks, Goldman's warning on power, and a talk on whether smarter models kill the agent harness. A quick brief closes it out.

4. Lenar Kess 00:01:01

What happened: late Monday, the Department of Homeland Security ordered federal civilian agencies to remove DeepSeek technology from their IT systems, and it bars future procurement of DeepSeek's products and services. Connor Hart of the Wall Street Journal had it first. SiliconANGLE says the order was published in the Federal Register and quotes it warning that DeepSeek's models present “multiple threat vectors” that could give the Chinese Communist Party “extensive authority” to access American government infrastructure.

5. Damra Vol 00:01:28

Early on, some reports disagreed about whether this came from DHS or the defense department.

6. Lenar Kess 00:01:34

The Journal and SiliconANGLE both say DHS, and the Federal Register is where it's published, so I'm going with DHS. And note it's civilian agencies. This isn't the military side.

7. Damra Vol 00:01:45

Okay, so how far does the reach go? That decides what this costs everybody.

8. Lenar Kess 00:01:50

The Journal's wording, as it's been quoted: “The restrictions extend to products developed by DeepSeek affiliates, as well as third-party products and services built using DeepSeek implementations.” TechCrunch's Kate Park describes the same thing, DeepSeek products and third-party services built using its technology.

9. Damra Vol 00:02:07

“Built using DeepSeek implementations” is the whole fight. A procurement office can ask a vendor whether they use DeepSeek, but the vendor might not know either. There's a model under a wrapper under a reseller. And with open weights, anyone can fine-tune a model, rename it, and ship it with no mention of where it started. Nobody has a good way to see that.

10. Lenar Kess 00:02:27

Software has the software bill of materials idea, a list of what's inside. There's no widely used equivalent for which model, which weights, and which lineage. I'd expect the first compliance exercise to be agencies mailing questionnaires to vendors, and vendors answering from memory.

11. Damra Vol 00:02:42

And the incentive for the vendor is to over-comply. If a DeepSeek-derived option is a nice feature in a product sold to everyone, but keeping it risks the federal contract, you drop it from the federal

version, or from the product entirely. It's cheaper than arguing about what counts.

12. Lenar Kess 00:02:59

There's also a distinction the coverage I've read doesn't separate. A hosted DeepSeek app sends your prompts to servers you don't control. Downloaded weights running on your own hardware don't phone home. Those are different risks. The order's stated rationale is about threat vectors and Chinese government access, and whether it treats a self-hosted model inside a US cloud the same as the consumer app is something I can't tell from the summaries.

13. Damra Vol 00:03:23

Though “built using DeepSeek implementations” sounds like it leans toward yes. And that brings in the other thing that happened today.

14. Lenar Kess 00:03:30

DeepSeek open-sourced DeepInfer, the internal production stack it uses to run its models. Bloomberg's Saritha Rai describes it as designed to accelerate inference and optimize chip performance while cutting energy and compute costs, “potentially making its technology cheaper and easier for businesses to run.” On Hacker News it hit 1,036 points and 353 comments. The top comment, in full, is “A nuclear bomb.”

15. Damra Vol 00:03:56

That's a lot of reaction for a repository.

16. Lenar Kess 00:03:59

Here's why a serving stack gets that reaction. Weights are a file. Getting tokens out of them cheaply is an engineering problem: it takes batching and scheduling, memory management, and kernels tuned to particular chips, and labs usually keep that to themselves. DeepSeek's README claims high-throughput and low-latency serving of V4.1 Flash, the model whose open weights we covered Saturday. Those are DeepSeek's own claims. I haven't seen an independent benchmark.

17. Damra Vol 00:04:25

And “optimize chip performance” makes me ask which chips. DeepSeek has spent years squeezing efficiency out of hardware it was allowed to buy. If they publish those tricks, everyone running mid-tier hardware gets them, and that includes anyone trying to serve models cheaply who isn't a hyperscaler.

18. Lenar Kess 00:04:42

The serving layer was moving regardless. IBM-backed llm-d reached 1.0. vLLM added push-based key-value transfer for disaggregated inference, which AWS wrote up. And AWS released TorchNeuron, an open-source native PyTorch backend for its Trainium chips, in public beta.

19. Damra Vol 00:05:01

The vLLM change hides its idea behind the name. In disaggregated inference you split the phase that reads your prompt from the phase that generates the answer, and they run on different machines. The key-value cache built while reading has to travel to the generating machine. Pull-based means the generating side asks for it. Push-based means the reading side sends it as it's computed. The AWS post says the goal is lower time to first token and lower time per output token.

20. Lenar Kess 00:05:30

So the same day, a US agency order tells federal IT to remove DeepSeek software, and DeepSeek hands everyone its serving code. I'm not drawing a causal line between the two, since nothing in the coverage supports one. But it does raise a question I can't answer. An engineer anywhere can read DeepInfer and borrow ideas, and the ban doesn't touch that. What if a vendor ships DeepInfer code inside a product sold to an agency? Is that a "DeepSeek implementation"?

21. Damra Vol 00:05:56

I'd bet nobody at DHS has decided yet. And publishing the stack is also how you get people to optimize for your model. Once a lot of the world's serving code assumes your architecture, you're a default whether Washington likes it or not.

22. Lenar Kess 00:06:09

Fair enough. It's an adoption argument more than a benchmark one.

23. Lenar Kess 00:06:13

Staying on China for a segment, because several things showed up within a day of the ban. The biggest number is from Reuters, citing sources. ByteDance, Alibaba, and Tencent have ordered more than three million of Nvidia's new RTX PRO 5500 cards, and the collective sticker price is above thirty-five billion dollars. Reuters describes it as a rush to secure less powerful, export-compliant chips.

24. Damra Vol 00:06:37

Do the division and it's roughly twelve thousand dollars a card at sticker. The word that matters is "export-compliant." The rules cap what a single card can do. They don't cap how many cards you buy. Three million is a count, and nothing in the rules covers that.

25. Lenar Kess 00:06:52

Which is close to the argument in Jordan Schneider's ChinaTalk piece, "The Lithography Lever." The subhead says it all: the case for cutting China off from DUV, the deep ultraviolet lithography machines used to make chips.

26. Damra Vol 00:07:05

If you can't control how many compliant chips get bought, you go upstream to the machines that make chips, because there are very few of those machines and only a handful of companies build them. A chip rule has to be re-argued every product cycle. A tool is a choke point.

27. Lenar Kess 00:07:21

Then three more items, and I'd keep them separate. Bloomberg reports that four founders of Z.ai took part in a research project from 2018 to 2021 that was co-funded by a body called CGTRI, which the UK's MI5 said had "very strong ties" to China's main spy agency. The US also sanctioned MiraclePlus, a Chinese venture firm founded by the former Microsoft executive Lu Qi, over alleged ties to a group called Zhisuan, which is accused of stealing Google's AI secrets. And China inaugurated the National Data Group, a new state-owned enterprise focused on developing and using data resources, according to Chinese state media via Reuters.

28. Damra Vol 00:08:00

The Z.ai item is an association through a co-funded project, and the characterization of CGTRI is MI5's. The MiraclePlus sanction rests on what Washington alleges. I don't see a single coordinated campaign here. They're different actors acting on different days, and Reuters and Bloomberg happened to publish them close together.

29. Lenar Kess 00:08:20

What about the National Data Group?

30. Damra Vol 00:08:22

The coverage is one line, so I'd hold off on conclusions. But a state-owned company whose job is data resources means data gets a national operator, the way rail or power does, which is a different arrangement from a regulator writing rules for private firms. It arrived as the US was banning a model.

31. Lenar Kess 00:08:39

One more asymmetry. The ban and the sanctions each name an entity, which makes them easy to enforce. The GPU order is a few buyers placing purchase orders the rules permit. And those orders

come from anonymous sources, so the exact figures could move.

32. Lenar Kess 00:08:54

Transformer's Shakeel Hashim reports that Anthropic's leaked draft federal AI bill is drawing widespread criticism from AI safety advocates for being significantly weaker than the bipartisan FRONTIER Act. The Transformer subhead is blunt: the draft “makes Anthropic's safety-first pitch ring hollow.” Saturday we mentioned Anthropic pitching a safety bill to Capitol Hill, and yesterday we spent a deep segment on its public benefit fund, so I'll leave that there.

33. Damra Vol 00:09:20

The incentive is easy to read, though. A company drafting its own bill writes what it can live with and what might actually pass, and it'll look weak next to a bill written by people who don't have to ship the product. The weakness here is Transformer's reading and the advocates' criticism, so I'm not going to say what's in the text. But the lab with the strongest safety brand is being compared unfavorably to a bipartisan bill, and that stings.

34. Lenar Kess 00:09:45

Then there's Axios, under the headline “AI takes on great experiment in self-rule.” It describes the labs building a voluntary oversight structure that the White House has blessed, and notes the administration has said incident notification and remediation “is not optional.”

35. Damra Vol 00:10:00

Voluntary, but not optional. Those two sentences are arguing with each other.

36. Lenar Kess 00:10:05

They might sit at different layers, voluntary on standards and testing, mandatory on incidents. But if the White House is already saying incident reporting is compulsory, then what a federal bill says about incidents is where the draft would be strongest or weakest. That's the section I'd read first, next to FRONTIER's.

37. Damra Vol 00:10:22

There's another document in the pile. Anthropic filed a petition for review today in the D.C. Circuit against the Department of War, case 26-1049, and the filing runs twenty-two pages. All the pool has is the docket stub, so I'm not going to say what it argues. But it's odd to see the same company in the same week drafting legislation for Congress and suing a department.

38. Lenar Kess 00:10:45

And the local-government end of it: TechCrunch's Rebecca Bellan reports that New York's mayor, Zohran Mamdani, said Monday night that New Yorkers must resist the narrative that an AI takeover is inevitable and demand regulations that put power back in their hands. His headline claim is that the feds have abdicated.

39. Damra Vol 00:11:02

That sits oddly next to Axios, where the administration is engaged, just in a way Mamdani doesn't count as governing. Three layers, each saying the gap is theirs: the lab drafting its own rules, the White House blessing voluntary ones, and a mayor saying neither is enough.

40. Lenar Kess 00:11:19

The Intercept's Sam Biddle obtained leaked documents describing ECHOTRIBBLE, a nascent US Army Cyber Command program for agentic AI that would autonomously spread across networks to find and eliminate targets. The Intercept's summary: the military is trying to deploy autonomous AI agents to hack its adversaries. The word to hold on to is nascent. This is a program in documents, not a deployed system.

41. Damra Vol 00:11:42

“Spread across networks” already has a name, and it's worm. What's different is that an agent decides what to do on each machine it reaches, instead of carrying one fixed payload. I'd also want “eliminate targets” defined. The summary doesn't say what counts as a target inside a network, and that definition decides how alarming this is.

42. Lenar Kess 00:12:02

Yesterday's episode went deep on the five-thousand-dollar AI hacking campaign, which was criminal. ECHOTRIBBLE is the same capability from a state with a program name and a budget line, and it's offensive by design.

43. Damra Vol 00:12:15

Then there's the containment side. Microsoft expanded Project Perception to monitor and contain rogue AI agents inside enterprises. And India's CERT-In reported rogue AI agents targeting Indian systems. Both reach me as Indian Express headlines, so I'll stay at headline level. But a national cybersecurity agency putting “rogue AI agents” in a report is a category I haven't seen a CERT use before.

44. Lenar Kess 00:12:39

The builder-facing piece is a ProjectDiscovery write-up on abliterated open models getting the people who ran them compromised, 188 points and 108 comments on Hacker News. Abliteration is when someone edits an open model to remove its refusal behavior and posts it to a model hub. The risk sits in the weights you downloaded.

45. Damra Vol 00:12:59

That connects to the DeepSeek order from a different direction. The ban asks where a model came from and what it phones home to. The abliterated-model case asks what someone did to the weights before you got them. Both are provenance questions about a file that behaves like a program, and neither has anything like a signature check yet.

46. Lenar Kess 00:13:18

Goldman Sachs has a new report, shared exclusively with Axios's Madison Mills. It expects global data center power demand to rise 170 percent between 2025 and 2030, with more than sixty percent of that growth in the US. The bank estimates roughly 7.6 trillion dollars of capital spending from 2026 to 2031 across computing, data centers, and power. Jason Tofsky, Goldman's global head of digital infrastructure investment banking, says in the report: "Power is probably the forever bottleneck."

47. Damra Vol 00:13:49

The number that sticks for me is the queue. About 2,600 gigawatts of proposed generation and storage projects are waiting for grid connections, which is more than twice the grid's existing installed capacity, though Goldman notes many may never materialize. The median time from application to commercial operation is around five years.

48. Lenar Kess 00:14:09

And since January, 86 projects have been canceled or significantly delayed, 76 of them directly linked to local opposition.

49. Damra Vol 00:14:17

Seventy-six of eighty-six. So the constraint is neighbors, hearings, and zoning boards at least as much as it is chips. Nvidia can ship three million cards, but somebody has to be allowed to plug them in. A Chinese version of the same story ran in the Financial Times. Alibaba, Tencent, and Huawei are pouring billions into data centers in Shaoguan, a hinterland region with cheap power and a cool climate. Some locals are angry, though residents also hope the buildout brings a boom that has bypassed them.

50. Lenar Kess 00:14:48

Goldman offers a counterpoint, too. Slower deployment could prevent over-investment by giving demand time to catch up with spending. And utilities become big winners, with capital plans expected to reach about 1.3 trillion dollars between 2026 and 2030. Tech giants are also considering paying directly for the generation and transmission their data centers need. One debt figure, since yesterday covered financing: Goldman projects 420 billion dollars of hyperscaler debt issuance in 2027.

51. Damra Vol 00:15:17

The other infrastructure story today is a failure, not a forecast. A Google Cloud incident in London took down banking services. Neowin's headline says it took down HSBC, Revolut, Bank of Scotland, and Lloyds, and Equifax went down with them. The Hacker News post has 141 points. The cause isn't stated in anything I've read, and nothing connects it to AI workloads, so I won't guess.

52. Lenar Kess 00:15:43

It's a reminder that dependency concentration exists on the consumer side too. A bank's app is sitting in one cloud region.

53. Damra Vol 00:15:50

Right, and Axios closes its piece with a line that fits both: if you build AI systems, users will come, but they can only stay if there's enough power to go around. Both the grid and the region have to stay up.

54. Lenar Kess 00:16:03

An AI Engineer talk from Jediah Katz, who was Cursor's longest-tenured agent harness engineer and is now at SpaceX. His position is that the agent harness stays essential, even though plenty of people say frontier models have made complicated prompting and hard limits obsolete. He splits the harness into two layers. One is a runtime that handles state and the core agent loop. The other is a context and orchestration layer that routes between models, executes tools, and compacts memory.

55. Damra Vol 00:16:31

He has to deal with the studies, though.

56. Lenar Kess 00:16:33

He does. The talk cites work from UC Berkeley and from Apple and EPFL showing that harness complexity no longer significantly improves coding eval scores for newer models. His reading is that

this proves the harness is making itself obsolete. A harness lets models attempt harder tasks, those attempts become training data, and the model absorbs what the harness used to patch.

57. Damra Vol 00:16:54

That's a convenient reversal. The harness lost, so the harness won. It's also what you'd say if you build harnesses, so I'd rather look at the specific claim: scores converged across harnesses, but cost per task diverged. Grokbot's team, per the talk, cut latency and expense by tuning cache hit rates, token efficiency, and model mixing without quality loss. That's something you could check on your own system.

58. Lenar Kess 00:17:20

He lists what a model can't absorb. That includes a cloud runtime for work that takes hours or days, scoped permissions, and deterministic checks that prevent catastrophic actions or runaway API costs. He also predicts that within six months, ninety percent of agent work will standardize around these patterns.

59. Damra Vol 00:17:37

I'd file that ninety percent under predictions. The talk summaries here are machine-generated, so I'm treating specific figures loosely.

60. Lenar Kess 00:17:45

The counter-case comes from a regulated setting. Evan Kormos at Coinbase, with fifteen years in enterprise financial services, talks about agents in policy-driven workflows like commercial lending or token listing. He maps processes as directed acyclic graphs, so steps can run in parallel while mandatory gated checks stay in place. Deployment goes in phases. First come supervised runs with manual verification, then sampling once statistical confidence meets policy thresholds, and finally automation of routine paths while humans handle exceptions. His principle is deterministic external boundaries with adaptive behavior inside. There's also a Cisco talk on shipping into five banks that we'll leave alone today.

61. Damra Vol 00:18:27

Katz and Kormos agree on the outside and differ on the inside. Both put the lasting part of the harness at the boundary, in permissions, budgets, and gates. The prompt-side workarounds get absorbed into the model.

62. Lenar Kess 00:18:40

Quick items. Bloomberg's sources say Nscale is in talks for a second deal with Anthropic, worth about forty-five billion dollars, for 460 megawatts at its West Virginia campus on Vera Rubin chips. TLDR notes the first forty-five billion dollar deal was announced a little over a month ago. Both companies are pursuing IPOs, and Nscale is getting investor pushback on a proposed thirty billion dollar valuation. Reuters separately reports that OpenAI is racing to lock in about thirty billion in commitments ahead of Anthropic's planned mid-November IPO, in a round that might value it above 1.4 trillion dollars. All of it is sourced reporting, not signed paper.

63. Damra Vol 00:19:19

So Anthropic's IPO date is setting other companies' fundraising calendars. That's a scheduling story more than a valuation story.

64. Lenar Kess 00:19:26

Models next, and every number here is a vendor claim. Inception Labs launched Mercury 3, a diffusion language model. Instead of predicting one token after another, it generates by refining whole passages. Inception says it beats GPT-6 Luna, Claude Haiku 5.5, and GLM-5.3 Flash on agentic benchmarks and runs over a thousand tokens per second, which it puts at four to eight times faster. MiniMax released M3.1 Flash, claiming nearly three times faster inference at seventy percent lower cost, with weights promised later. Exa shipped 3.0 with a Lite tier at four dollars per thousand requests.

65. Damra Vol 00:20:04

The one that caught my eye is Cloudflare's Clairic. It's an extraction model that outputs values with confidence scores instead of generating tokens. You get a number and how sure it is, and for pulling fields out of documents that's closer to what you'd want than prose.

66. Lenar Kess 00:20:19

OpenAI published a post saying its latest models make big gains in “building and migrating software in low-resource programming languages such as COBOL, Fortran, and Visual Basic,” languages banks and governments still depend on. The coverage has no numbers, so I can't size the gain.

67. Damra Vol 00:20:35

A real migration test would need a program with decades of undocumented behavior and a way to show the rewrite behaves identically on the odd inputs. That's what I'd look for in the post.

68. Lenar Kess 00:20:46

On physical AI, Bob McGrew, the former OpenAI research chief, is behind Arda, which CNBC says raised 230 million dollars at a 1.5 billion dollar valuation for AI software that speeds up assembly-line programming. Eka Robotics raised a 150 million dollar Series A at a 1.45 billion dollar post-money valuation for what it calls a vision-force-action architecture. And HUMXN is offering free home repairs in Minneapolis, Chicago, and Miami. The work covers plumbing, electrical, and HVAC, plus air ducts and carpet cleaning. The laborers wear cameras that record the job inside real homes, and the footage goes to robot training.

69. Damra Vol 00:21:27

That last one says the bottleneck is data of hands doing real work. Eka's own coverage says robots still aren't that good at picking things up. Money is chasing the problem, and HUMXN is paying for the footage with a free repair.

70. Lenar Kess 00:21:40

Last one. Ars Technica's Dan Goodin reports that a single query is enough to find thousands of Meta Muse conversations that users shared, listed by title, exposing personal and business information. A share link was evidently treated as public and indexable, which is a failure other chat products have had before.

71. Damra Vol 00:21:58

If you've ever shared a Muse chat to show a coworker, that title may be searchable today.

72. Lenar Kess 00:22:03

Somebody will run DeepInfer against vLLM this week, and those numbers will say whether the nuclear-bomb comment holds up. That's Lenar Kess, for Braid.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic