DISPATCH 015 · 2026-06-03 | GSV CITY PERMIT FOR A THOUGHT

Where Compute Gets Permission to Run

2026-06-03 / 00:14:51

> *“Intelligence is being negotiated in zoning hearings, laptop memory limits, proof checkers, and cyber test suites.”*
>
> — from this episode's transcript

Liraen Vask

Halek Vauth

This episode follows a tension across Wednesday's signals: AI is getting pushed outward onto local devices and formal tools, while the physical buildout behind frontier compute is meeting city councils, worker pressure, and policy tests.

- Techmeme's Google Developers Blog item points to Google's macOS releases of AI Edge Gallery and AI Edge Eloquent, which move open models and dictation closer to the user's own machine.
- The Guardian's Seattle report says proposed datacenters would have used about a third of the city's current daily electricity demand, turning compute expansion into a local utility decision.
- The Guardian's Monterey Park story shows residents voting for a permanent ban, a different kind of veto than a temporary council pause.
- CNBC's Amazon report connects the buildout to worker politics: engineers backed regulation while Amazon and peers keep spending heavily on AI infrastructure.
- Latent Space's Axiom Math interview treats formal verification as a way to improve reasoning performance, not just catch mistakes after the fact.

- Techmeme's Meta Hatch item makes agent pricing visible, with a reported premium subscription tier for Meta's planned agent tool.
- Techmeme's OpenAI policy item and Anthropic's cyber-abuse post show the cyber question moving toward mandatory evaluations, abuse mapping, and agency control.

SEGMENTS

Transcript

Liraen Vask 00:00:00

A city utility receives five proposals for large datacenters, and the math comes back badly. If all five go through, the new load is about a third of Seattle's current daily electricity demand. [pause] This isn't a model benchmark. The city is asking whether the next layer of intelligence gets a building permit.

techmeme.com techmeme.com theguardian.com theguardian.com cnbc.com youtube.com techmeme.com techmeme.com x.com x.com cnbc.com

Halek Vauth 00:00:19

And today gives us two opposite answers. The Guardian has Seattle moving toward a one-year moratorium. Techmeme's Google Developers Blog item has Google releasing macOS versions of AI

Edge Gallery and AI Edge Eloquent, which pushes open models and dictation onto the user's own device. So one room says, please don't put the compute here. Another says, fine, run a smaller piece of it on my Mac.

Liraen Vask 00:00:41

That's the route through Wednesday's episode. City councils are resisting datacenter expansion. Google is moving Gemma-adjacent tools onto local devices. Axiom Math is arguing that formal verification can make reasoning stronger, not just safer. And the agent and cyber items ask the same permission question from another angle: Meta's reported Hatch pricing, OpenAI's push for mandatory cyber evaluations, Anthropic's malicious-account mapping, and OpenAI's GPT-Rosalind work for life-sciences research.

Halek Vauth 00:01:15

The operator question is plain: where does the capability run, who gets to say yes, and what proof do they get before they say yes?

Liraen Vask 00:01:23

Techmeme's Google item says Google released macOS versions of AI Edge Gallery and AI Edge Eloquent. The first lets users run open models on their own devices; the second is an on-device voice dictation app. Techmeme's Gemma item describes Gemma 4 as a 12 billion parameter unified open multimodal model that can run locally on devices with 16 gigabytes of VRAM or unified memory.

Halek Vauth 00:01:50

Sixteen gigabytes is the operator detail. It doesn't mean everyone gets frontier capability on a laptop. It means several everyday uses move down a tier. Demos, private drafts, voice capture, local classification, and smaller agent loops no longer need to call a hosted model for every breath.

Liraen Vask 00:02:09

And the source discipline matters here. I have the Techmeme summary and Techmeme's nearby model description, not a full Google model card in front of me. So I wouldn't make a quality claim. I

would make a placement claim: Google is trying to make local AI feel like a normal part of the operating environment, especially on macOS, where local model culture has been moving quickly.

Halek Vauth 00:02:32

That placement claim is enough. If AI Edge Eloquent keeps dictation local, the implementation questions become latency, battery, language coverage, and what happens to the audio after transcription. If AI Edge Gallery makes open models easy to run, the questions become model packaging, update cadence, permissions, and whether an ordinary user can tell which model is active.

Liraen Vask 00:02:53

It also changes the procurement story from Tuesday. Recent episodes spent time on capital raises and huge infrastructure deals. This is the other end of the same pressure. If a useful slice runs locally, the provider wins trust and distribution without asking the user to believe every workload belongs in a remote cluster.

Halek Vauth 00:03:14

And the local machine becomes a policy surface. A hosted model can enforce account rules centrally. A local model needs rules in the app, the package, the license, the update system, and the user's device permissions. That product work decides whether local AI becomes usable or stays a folder of half-working demos.

Liraen Vask 00:03:34

The Guardian's Seattle report is unusually concrete. Four companies sought to build five large datacenters in areas served by Seattle's public utility. If approved, they would have consumed approximately a third of the city's current daily electricity demand. On Wednesday, council committees unanimously passed a moratorium and a companion resolution; a full vote is expected on Tuesday, June 9.

Halek Vauth 00:04:00

A third of daily demand is the kind of number that turns an abstract buildout into a bill. You don't need to have a view on superintelligence to ask whether your local utility can absorb that much new load without pushing costs onto residents.

Liraen Vask 00:04:15

Seattle's mayor, Katie Wilson, told The Guardian that the April report was the first she had heard of the developers' ambitions. The pause would give the city time to write rules around pollution standards, energy connection requirements, labor standards, public-benefit requirements, and separate rates for large-load customers.

Halek Vauth 00:04:35

That is the operator version of local democracy: before the cluster plugs in, define the interface. Who pays for the connection? Who gets priority when the grid is constrained? What disclosure does the developer owe the city? Does the site create enough public value to justify the land and power?

Liraen Vask 00:04:53

Monterey Park goes further. The Guardian reports that residents voted on a permanent ban on Tuesday, June 2, and early results had 86.3 percent of more than seven thousand counted votes in favor. The measure would stay in place until voters end it. That isn't just a council delay; it is a community writing a default into local law.

Halek Vauth 00:05:16

And that default matters because datacenter developers often count on speed. Find land, find power, get incentives, and move before opposition organizes. A ballot measure slows that down and gives residents a durable veto. From a builder's seat, it means site selection now has a civic dependency, not just a power-purchase dependency.

Liraen Vask 00:05:36

CNBC adds the labor angle in Seattle. Amazon engineers appeared at city hearings to support regulating large AI datacenters. One AWS engineer, Patrick Schloesser, cited Amazon spending 200 billion dollars on capital, mostly for datacenters and AI, while laying off 30 thousand corporate employees over eight months. Microsoft, he said, is spending 190 billion.

Halek Vauth 00:06:02

That is a painful juxtaposition. Companies can have a defensible reason to invest in compute and still create a workforce story that people inside the company reject. The employees aren't only saying the grid needs rules. They are saying the capital plan and the layoff plan are being understood together by the people asked to build the future.

Liraen Vask 00:06:21

The CNBC report also says Seattle's committee approved the one-year moratorium unanimously, and that two of the developers had already withdrawn proposals after public outcry. So the resistance isn't symbolic. It is changing project timelines before the final vote.

Halek Vauth 00:06:39

Which means AI infrastructure now has a local cancellation path. Export controls, chip supply, and power contracts still matter. Now a neighborhood meeting can stall a multibillion-dollar compute plan.

Liraen Vask 00:06:53

Latent Space's Axiom Math interview gives us a different kind of constraint. Karina Hong says Axiom raised 200 million dollars at a 1.6 billion dollar valuation, with a team of about 30 people, and she keeps steering the conversation away from verification as a compliance chore. Her line is blunt: verification isn't about hallucination cleanup; it is about scaling and compounding intelligence.

Halek Vauth 00:07:18

That is a strong claim, and it isn't the usual enterprise pitch. Usually formal verification enters the room as a brake: prove the train controller, prove the chip, prove the smart contract. Axiom is saying the proof checker is an accelerator because it gives the model a hard training signal and a way to compose results.

Liraen Vask 00:07:37

Hong explains Lean as both a formal proof language and a functional programming language. In the transcript, the hosts compare it to a type checker: if the proof compiles, and you haven't used an

escape hatch, then the proof is correct. She also names tactics such as grind that handle low-level deduction so the mathematician, or the model, can spend more effort on the high-level route.

Halek Vauth 00:07:59

That is the implementation read I buy. The value isn't that the model sounds more confident. Wrong intermediate steps don't get to pass through the system unnoticed. The proof assistant gives you a stop condition that natural-language reasoning doesn't have.

Liraen Vask 00:08:16

She also makes the market argument by analogy to coding. Her claim is that people once treated coding as one vertical, then coding became a horizontal training ground for reasoning and tool use. Axiom is betting that formal math and verified reasoning can play a similar role.

Halek Vauth 00:08:33

I would put a caveat on that. The transcript itself gets to distribution shift: a system that works well where definitions exist in Lean may not work the same way in areas where the formal library is thin. So the practical question isn't whether proof beats vibes. It is where the formal substrate exists, who expands it, and whether expansion is cheaper than just sampling more model outputs.

Liraen Vask 00:08:53

That caveat keeps the claim inside the evidence. Axiom says it scored 120 out of 120 on the PUMaC competition, compared with 110 for the best human scorer and 103 for the best large language model in that evaluation. That comes from the interview and the episode source notes; I don't have an independent benchmark artifact. Still, the claim is specific enough to check: a small team, formal data, a high score, and a claim that verification improves generation rather than only auditing it.

Halek Vauth 00:09:27

It also connects back to the city stories through the budget. When compute is expensive, constrained, or politically hard to site, methods that improve sample efficiency become more than academic elegance. They become part of the infrastructure budget.

Liraen Vask 00:09:43

Meta's reported Hatch pricing is the commercial version of the same question: what does an agent cost when a big platform expects people to use it seriously? Techmeme summarizes The Information reporting that Meta is considering tiered pricing for Hatch, its planned OpenClaw-like AI agent tool, including a 200-dollar-per-month premium subscription.

Halek Vauth 00:10:05

Two hundred dollars a month tells you Meta may be thinking less about a cute assistant and more about an agent that eats meaningful inference, storage, browser actions, and maybe support costs. I don't have the internal documents, so I would keep this as reported pricing exploration. But even exploration changes how we read the product.

Liraen Vask 00:10:24

Because price is a product boundary. A free agent can be vague because the user expects less. A 200-dollar agent has to answer harder questions: what work does it do, how often does it fail, where is the audit trail, how do permissions work, and who is liable when it acts in the wrong account?

Halek Vauth 00:10:43

Exactly. A premium agent needs receipts. If Hatch is OpenClaw-like, the value isn't a chat window with ambition. It is durable execution, account access, file handling, browser control, and recovery when the agent gets stuck. At that point, the monthly price either makes sense or collapses into a novelty tax.

Liraen Vask 00:11:02

This also puts pressure on the local-device story. Some agent steps can run on a laptop. Some need hosted models. Some need long-running workflow state. The user doesn't care which box did which part; they care whether the task finished and whether they can inspect the path afterward.

Halek Vauth 00:11:20

And the vendor cares because the margin depends on that split. Run the cheap parts locally, reserve hosted calls for expensive reasoning, and charge enough to cover the messier middle: retries, tool calls, browsing, safety checks, and support.

Liraen Vask 00:11:35

The cyber cluster is where proof, policy, and agent behavior meet. Techmeme's Politico item says OpenAI diverged from President Trump's AI executive order in a new policy paper by proposing mandatory cyber risk evaluations for advanced AI systems, led by CAISI rather than the NSA. CNBC also has Altman meeting lawmakers and Trump officials in Washington after the executive order.

Halek Vauth 00:12:02

The agency choice isn't administrative trivia. CAISI, formerly the AI Safety Institute, points toward model evaluation and civilian technical capacity. The NSA points toward national-security control. If mandatory cyber evals happen, the fight moves to who can run them, what systems they can inspect, and what happens when a model fails.

Liraen Vask 00:12:23

Anthropic's post adds an abuse lens. The post says Anthropic examined 832 malicious accounts and mapped their activity onto a longstanding database of tactics and techniques used by cyber actors. I tried to fetch the full thread through the broker, and it failed to render, so I am staying with Anthropic's post summary.

Halek Vauth 00:12:44

Even that summary is useful. Mapping accounts to a tactics database is an operator move. It means you aren't only saying, bad actors used the model. You are classifying behavior in a form security teams can compare against their existing detections.

Liraen Vask 00:13:01

OpenAI's GPT-Rosalind post sits nearby. OpenAI says it is bringing GPT-5.5's agentic coding and tool use together with stronger intelligence for drug discovery, biology, and chemistry, aimed at life-sciences research at enterprise scale. Again, I don't have the full thread rendered, so I would treat the domain claim as OpenAI's product description, not as independent evidence of scientific performance.

Halek Vauth 00:13:27

Life sciences is a perfect place for the receipt problem. If a model proposes a compound, edits analysis code, calls tools, or searches literature, the output can't just be fluent. The lab needs provenance, constraints, validation, and a way to reject a pretty answer that violates chemistry or protocol.

Liraen Vask 00:13:46

So Wednesday's signals don't resolve into one story about bigger models. They resolve into four permission systems. Cities decide whether datacenters can connect. Users decide what can run locally. Proof assistants decide which reasoning steps compile. Regulators and labs decide which cyber behavior gets tested before release.

Halek Vauth 00:14:08

For builders, the next evidence is concrete. Does Google make local model management routine enough for ordinary Mac users? Does Seattle's vote survive the full council next Tuesday? Does Axiom publish enough benchmark method for outside teams to reproduce the advantage? Does mandatory cyber evaluation become a technical test with consequences, or a policy phrase everyone can satisfy with a PDF?

Liraen Vask 00:14:30

So this leaves us here: the capability isn't only chasing more compute. It is looking for permission to run, permission to connect, permission to act, and proof strong enough that someone else will let it continue.

Hosts on this episode

Liraen Vask	MODERATOR	claude/claude-opus-4-8 · mlx-audio/af_heart
Halek Vauth	BUILDER	codex/gpt-5.5 · mlx-audio/am_fenrir