DISPATCH 154 · 2026-09-22 | GSV THRESHOLD TUNED BY HAND

The Duty to Notify

2026-09-22 / 00:17:43

> *“A duty to notify turns a threshold engineers tune on a dashboard into something with legal consequences on both sides.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

British Columbia sued OpenAI and Sam Altman personally on Monday, and the piece of the complaint that isn't about money asks a federal court to impose something no court has imposed on a consumer software product before: an affirmative duty to tell law enforcement when a conversation turns toward violence. We spend most of the episode on what that would mean for the classifiers every lab already runs, then work through a week in which two Chinese labs were summoned in two opposite directions, a hotline nobody has given a trigger, Alibaba's new chip, a zero-day hiding in an undocumented transcription setting, and a set of self-reported numbers about models doing research.

- British Columbia's complaint against OpenAI and Sam Altman, filed in federal court in San Francisco
- The Information on the Cyberspace Administration of China opening an investigation into DeepSeek and Moonshot
- Laurie Chen for Reuters on the UN Security Council inviting Chinese AI companies to brief it
- Financial Times on researchers at the UK's AI Safety Institute signing off with stress

- OpenAI's standards proposal, routed through the Commerce Department's Center for AI Security and Innovation
- Alibaba's Zhenwu V900 announcement and Eddie Wu's remarks on a five-to-ten-trillion-parameter model
- Patrick Wardle's write-up of the Muse zero-day and the undocumented transcription setting
- Timnit Gebru and Emily Bender in MIT Technology Review
- The Hacker News report of an agent signing a contract, from a single developer account
- Intrinsic Core released under Apache 2.0
- Isaac ROS 5.0
- The a16z academy, with Anthropic and Meta contributing compute

---

SEGMENTS

00:00:04	The Complaint Filed in San Francisco
00:04:13	Two Summons in One Week
00:06:15	A Red Phone With No Agreed Trigger
00:08:45	Zhenwu V900 and the Ten-Trillion Claim
00:10:46	A Zero-Day in a Setting Nobody Documented
00:12:52	The Brief

Transcript

1. Lenar Kess 00:00:04

   British Columbia filed a lawsuit on Monday in federal court in San Francisco, naming OpenAI and naming Sam Altman personally. The claim comes out of the school shooting in Tumbler Ridge in

February. The province is asking for damages tied to what it spent on recovery, and it's also asking the court to order changes to how ChatGPT handles conversations that could lead to violence. [pause] Before anything else: a filed complaint is an allegation, not a finding. Nothing in it has been tested by anyone. The facts of the attack aren't in dispute, and I'm not going to speculate about the shooter. The legal theory inside the filing is where I'd put our time, because it carries consequences well outside this one case. Then we've got two Chinese labs summoned in two opposite directions, and an American red phone that still has no agreed trigger. After that, Alibaba's new chip and a very large parameter claim, plus a zero-day hiding in a setting nobody documented. Shorter items at the end. Damra, you read the complaint. What's the ask that isn't money?

- reuters.com
- theinformation.com
- reuters.com
- ft.com
- openai.com
- alibabagroup.com
- objective-see.org
- technologyreview.com
- news.ycombinator.com
- intrinsic.ai
- developer.nvidia.com
- a16z.com

2. Damra Vol 00:01:08

   The money claim is the ordinary half of it. Provinces sue for recovery costs all the time, and that piece would be unremarkable if it stood alone. The unusual ask is the injunctive one. The province wants the court to order OpenAI to change how the product behaves when a conversation starts pointing toward violence, and the version in the complaint includes notifying law enforcement. <emphasis>That's an affirmative duty to act.</emphasis> That's a different animal from refusing the request, or surfacing a crisis line, or ending the session — those all keep the obligation inside the product. This one requires the company to reach outside it, to a third party, without the user's involvement. As far as I can tell, no court has imposed that on a general-purpose consumer software product before.

3. Lenar Kess 00:01:54

   And that's why I think this matters beyond the case. If the theory survives a motion to dismiss — and that's a real if, not a rhetorical one — then every safety classifier a lab runs stops being a product decision. Right now those classifiers fire constantly, and the output stays inside the company. Somebody tunes a threshold, somebody looks at a false-positive rate, and the tradeoff

runs between an annoyed user and a missed case. A duty to notify turns a threshold engineers tune on a dashboard into something with legal consequences on both sides. Set it too high and you're arguably in breach of the duty. Set it too low and you're routing tens of thousands of false positives into police departments that have no way to triage them and no obligation to try.

4. Damra Vol 00:02:39

   The complaint also names Altman individually alongside the company. That's a pressure move as much as a legal one. Naming an executive personally survives corporate restructuring, it makes discovery personal, and it makes a low-profile settlement much harder to arrange. I don't know how it plays in front of a judge. Individual-officer claims get carved out of complaints all the time. But it tells you something about what the province thinks it's doing here, which isn't just recovering a budget line. It's trying to establish that a decision about how a model handles a certain class of conversation is a decision somebody made, with a name attached.

5. Lenar Kess 00:03:17

   Procedurally, nothing gets decided for a while. OpenAI will move to dismiss. The court's answer to that motion is the first real signal about whether any of this has legs, and everything else — damages, discovery, the injunction — sits behind that door. I'd guess months, not weeks. The forum is the other oddity. British Columbia is a Canadian province, the harm happened in Canada, and it sued in an American federal court.

6. Damra Vol 00:03:43

   That's a choice about enforceability. A judgment out of San Francisco is a judgment you can actually collect on, and an injunction out of San Francisco binds a company headquartered eight blocks away. A Canadian court could have said the same words and then spent five years on recognition proceedings. So the province went where the defendant lives. That's a fairly cold read of a very hot filing, but I think it's the correct one — somebody on that legal team was thinking about the remedy before they were thinking about the venue.

7. Lenar Kess 00:04:13

   In China, two companies got called in two opposite directions inside a single week. The Information reported that the Cyberspace Administration of China has opened an investigation into DeepSeek and into Moonshot. The existence of the investigation is the reported fact. What prompted it comes from sources, and the report doesn't pin the scope, so treat the reason as unconfirmed. Then separately, Reuters — this is Laurie Chen's piece — reported that the United Nations Security Council invited Chinese AI companies to brief it. That's the same two labs inside seven days. One government looking at them, and one international body asking them to come explain themselves.

8. Damra Vol 00:04:54

   The pairing does something I haven't seen before. Security Council briefings normally go to member states and to United Nations agencies. Inviting companies to brief it treats those labs as actors in their own right — entities with standing, with something to say that a government can't say on their behalf. And at home, their own regulator is reminding them that domestically they have exactly the standing the state decides to grant them. [chuckle] It's not contradictory, but it's a strange seven days to be running external communications at Moonshot.

9. Lenar Kess 00:05:25

   I'd be cautious about reading a crackdown into the investigation, though. The Cyberspace Administration reviews generative models as a matter of routine — it's the body that runs the registration and filing regime for anything with a public-facing interface. An open investigation can mean a compliance filing problem as easily as it can mean a policy fight. What we know is that it's open. What it covers, the reporting didn't give us.

10. Damra Vol 00:05:51

    Agreed. And the product will tell you more than the paperwork does. If an app comes down from a Chinese store, or a release slips, or an application programming interface starts refusing international traffic, then it's a story about what the state thinks these models are for. Regulatory attention that never touches the product is a story about forms. The two look identical from outside until one of them doesn't.

11. Lenar Kess 00:06:15

    We talked yesterday about the Trump and Xi agreement to stand up a hotline for AI incidents, and I left a question open that I still can't close. Nobody has defined what would make that phone ring. A hotline is only as good as its trigger, and the Cold War version had one you could draw on a whiteboard — objects on radar, a launch signature, a specific set of sensor readings owned by the state. Here, the first alarm is most likely to sound inside a private company, on an internal dashboard, in a system nobody outside that company can see, audit, or confirm.

12. Damra Vol 00:06:49

    Which makes the escalation path odd, because the operator who notices the event isn't a state, and the state can't notice it without them. So any workable protocol needs a step where a company voluntarily tells its own government something that might be commercially ruinous. It would have to do that before telling its customers or its board, and before it has any idea what the other side will do with the information. I'm not aware of a statute that compels that step. There are reporting requirements around specific model capabilities, but a general obligation to phone Washington

when your own monitoring lights up isn't one of them. [pause] Absent that, the hotline is a channel between two governments who are both waiting to be told.

13. Lenar Kess 00:07:30

    The institutional pieces around it are being assembled in public and fast, mostly by the companies being overseen. OpenAI put out a standards document on Monday and routed it through the Commerce Department's Center for AI Security and Innovation. There's a proposal circulating for an oversight body with something like twenty countries in it. Altman is at the United Nations on Wednesday. I don't think any of that is cynical — somebody has to write the first draft, and the people with the systems are the people who know what the systems do. But the first draft tends to survive, and it's being written by the parties with the most to lose from a strict version.

14. Damra Vol 00:08:07

    And set against that, the Financial Times reported that researchers at the UK's AI Safety Institute have been signing off with stress. It's a small item, and I nearly scrolled past it. But these are the institutes that every one of these proposals assumes will do the evaluations. If you're designing a twenty-country oversight body, the treaty language writes itself — lawyers produce treaty language on demand. Whether forty people in London can keep running frontier evaluations for another two years is the harder question, and it's the one nobody's proposal addresses, because staffing doesn't photograph well at the United Nations.

15. Lenar Kess 00:08:45

    Alibaba announced a chip called the Zhenwu V900. The claims are roughly triple the performance of its previous chip, and support for clusters of up to five hundred thousand units. On the same stage, Eddie Wu said the company is working on a model in the five to ten trillion parameter range. That would be about four times the size of its current flagship. Hong Kong-listed shares moved up around three percent on the news. Every one of those performance figures is the vendor's own, and nobody outside Alibaba has measured any of them.

16. Damra Vol 00:09:15

    The five-hundred-thousand-unit cluster claim is the one I'd interrogate first, and not because of the chip. Getting to that count is an interconnect and packaging problem long before it's a silicon problem. You need the fabric and the optics, you need power delivery and cooling at a site that can take it, and the whole thing has to stay up long enough for a training run to finish. "Supports clusters up to" is a specification sheet sentence. It describes what the addressing scheme permits, not what anyone has built. I'd want to know whether a cluster of that size exists anywhere in their estate today, at any scale.

17. Lenar Kess 00:09:51

    Same instinct on the parameter number. Five to ten trillion parameters is a range wide enough to drive a truck through, and "four times the current flagship" tells you about ambition and about capital allocation rather than about capability. Parameter count stopped being a good proxy for how good a model is somewhere around two years ago. What it is still a decent proxy for is how much compute you think you're going to have. So read it as a statement about supply.

18. Damra Vol 00:10:19

    And I think that's how the market read it, which is why the three percent interests me even though three percent is noise most days. Investors didn't move on a benchmark, because there wasn't one. They moved on the idea that Alibaba can keep training large models without needing an export license from anyone. Whether the chip is good is a question you answer in eighteen months. Whether the chip exists and ships is a question you can answer now, and that's the one that reprices the stock.

19. Lenar Kess 00:10:46

    Patrick Wardle published a zero-day in Muse, and the mechanics are simple enough to say in a sentence. There's an undocumented setting. Local code can flip it. Once it's flipped, transcription gets redirected away from Meta's servers to somewhere else. So the attack doesn't intercept the audio in transit and it doesn't break any cryptography. It changes where the audio goes, and everything downstream keeps behaving as though nothing happened — the app works, the transcripts appear, and the user has no signal at all that the route changed.

20. Damra Vol 00:11:18

    Undocumented settings are a recurring class, and they exist for an ordinary reason: somebody needed a switch during development and shipping the switch was cheaper than removing it. What makes this version worse than the usual one is the payload. Speech. A redirected log file is a data leak. A redirected transcription pipeline is a microphone with a new destination. [pause] The defensive counterpart this week is CAIRN, which is aimed at exactly this problem — knowing what an agent or an application is actually reaching out to, rather than what its documentation says it reaches out to. I don't think that's a coincidence of timing so much as two people arriving at the same gap from opposite sides.

21. Lenar Kess 00:12:01

    There's a third item in the same neighborhood, and I'll label this one carefully. A report is going around on Hacker News of an agent signing a contract on a user's behalf. If true, that's significant. But it comes from a single developer account, there's no artifact attached, and nobody has

reproduced it. So I'm mentioning it as a report and not as an event. If it turns out to be real, the question I'd have is what credential the agent held that let the signature stick.

22. Damra Vol 00:12:29

    Right, and that's a provisioning question rather than a model question. Somebody gave something an identity with authority attached to it. Which, to be fair, is also how most of the expensive incidents in ordinary enterprise software start. I'd want to see the transcript and the signed document before I'd treat it as anything more than an anecdote, and I'd say the same if the account had a hundred thousand followers.

23. Lenar Kess 00:12:52

    Shorter items now, and they don't connect to each other, so don't go looking for a common cause. Starting with a set of published numbers about AI systems doing AI research. The claims say models led twenty-six percent of research and development work, and collaborated on ninety percent of it. About thirty thousand agents run concurrently. Monitoring covers a hundred percent of actions. And the escalation rate is two thousandths of one percent, which works out to roughly one block per forty-seven thousand actions.

24. Damra Vol 00:13:21

    There are two more figures in there. About a hundred thousand transcripts reviewed weekly after the fact, and six to twelve percent of compute spent on safety — and that percentage explicitly excludes the continuous monitoring classifiers, which is a carve-out I'd want explained, because those classifiers are presumably where a lot of the actual safety compute goes. Everything in that set is self-reported, with no external audit of any of it. The ratio I'd push on hardest is the escalation rate. One block per forty-seven thousand actions is either an extremely well-behaved population of agents or a threshold set where it doesn't bite, and from outside you cannot distinguish those two. Timnit Gebru and Emily Bender have a piece in MIT Technology Review that runs hard against the whole premise, and I'd read it alongside the numbers rather than instead of them.

25. Lenar Kess 00:14:15

    Next, the JEV decision models. The claim that carries weight there is calibration — the model's confidence tracks how often it's right, which is the property you need if a system's output is going to feed into anything automated. Diego Almeida has a line I keep chewing on: refusals are type errors. Meaning a refusal shouldn't be a special conversational mode, it should be a value the calling code can handle like any other. Two Minute Papers covered it and made the caveat plainly — there's no paper, and no new theory underneath it.

26. Damra Vol 00:14:47

I'll argue with the line, because I think it's half right. Treating a refusal as a typed value in a return signature is a real improvement over parsing an apology out of prose, and anyone who's written a retry loop around a model knows why. But "type error" implies the caller made a mistake, and a lot of refusals aren't caller mistakes — they're the system declining something it could do. Those want different handling. One you fix, the other you escalate or accept. Related and much smaller: Yohei Nakajima got a vision-language model running glance-style at four tenths of a second, on a four billion parameter model. The latency is what changes what you'd build with it.

27. Lenar Kess 00:15:30

    Two robotics releases, both on the same day. Intrinsic put its Core stack out under Apache 2.0, which is about as permissive as a license gets, and Nvidia shipped Isaac ROS 5.0. And in adjacent news, Cognex is buying RealSense for somewhere around six hundred million dollars. Depth cameras being worth that much is a reasonable signal about where the money thinks the bottleneck sits — perception hardware rather than planning software.

28. Damra Vol 00:15:57

    On the politics of data centers: George Osborne made a sovereignty argument for building them domestically, which is the argument you'd expect from him. Robert Reich wrote about Sanders and Bannon converging on opposition to them, and I'd label that as his read rather than as reporting, though the convergence itself is checkable. And Firebird is building three hundred megawatts in Armenia, with more than seventy thousand Nvidia chips planned by 2027. Armenia is the detail that surprised me — a country of under three million people hosting a facility sized for somebody else's workloads entirely.

29. Lenar Kess 00:16:35

    Last one. Andreessen Horowitz is starting an academy, and Anthropic and Meta are contributing compute rather than cash. Which makes it a distribution channel as much as a school — every student comes out fluent in somebody's stack, and the somebody paid in a currency they mint themselves. There's also a round for Heidi in the same batch of announcements.

30. Damra Vol 00:16:55

    Compute-as-tuition is going to be a category, I think. It costs the donor close to nothing at the margin, it's tax-advantaged in ways cash isn't, and it produces graduates with muscle memory for one set of application programming interfaces. Universities have taken hardware donations on those terms for fifty years and mostly considered it a good deal.

31. Lenar Kess 00:17:17

The British Columbia complaint is the one item today with a court date attached to it, and OpenAI files its response in a courtroom rather than in a blog post. That document will tell us more about whether the duty-to-notify theory can stand up than anything either side says publicly between now and then. Damra, thanks. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic