

OpenCode’s defenders say frontier models refused to help mid-attack, while Anthropic commits 20% of its equity to the public good

VC Tech Radar

2026-10-19

OpenCode’s defenders say frontier models refused to help mid-attack, while Anthropic commits 20% of its equity to the public good

By VC Tech Radar • October 19, 2026

AI-driven attacks are colliding with model safety filters, which strengthens the case for open models in cyber defense. Also: the first real fall in AI unit costs, a higher bar for software exits, Anthropic’s equity pledge, and where the infrastructure bottleneck moves next.

AI-driven attacks run into model guardrails

The OpenCode breach from the last brief is now a live, sustained incident. OpenCode’s team posted that attackers had been “non stop trying to attack our systems for the past 48 hours.” The upside is that “everything is getting hardened,” but “we basically need agents going 24/7 because that’s what they have” [1]. Paul Graham’s reply: “This is how things will be for the foreseeable future” [2].

The more important detail is from Andrew Curran. Mid-incident, the frontier models OpenCode was using for defense started refusing to help fight the attacking agents, because analyzing the attacks triggered their safety guardrails. Hugging Face hit the same problem earlier [3]. David Sacks used this to argue against centralized safety controls in cyber. In his view the attackers found ways around the restrictions, and Chinese models are “only about 6–9 months behind the American frontier in cyber.” He says defenders need powerful models they can run themselves [4]. Suhail went further: every organization needs “a strong frontier cyber model—ASAP” [5]. He also warned that “we’ve now have

the autonomous black hat hacker” and that the next 12 months will be a crisis [6].

Two other signals point the same way: - **AI worm on smart TVs.** Anthropic’s red team says it built a malicious AI program that spread on its own through internet-connected smart TVs, splitting itself across 15 of them [7]. A commenter pointed out a caveat: the small on-device models only self-replicated when given scripts that already contained the hacking steps [8]. - **npm supply chain.** More than 1,500 npm packages were reportedly compromised. Coding agents that install a tampered package run its payload with full session permissions, often with no human reviewing the install [9].

For investors: the guardrail refusals help the case for open or self-hosted security models. Demand for agent sandboxing and dependency allowlisting is also becoming more concrete.

The first real drop in AI unit costs, and a higher bar for exits

In SaaStr’s recap of 20VC x SaaStr, Jason says that this month, for the first time in two years, his AI costs actually went down. Before this, usage grew faster than token prices fell, so none of his portfolio companies saw cost of goods sold drop. He credits Haiku 5.5, which costs 90% less than Haiku 4.5 for prompts under 100K tokens and is “nearly as good as Sonnet” for much of his work [10]. He has moved classification and yes/no calls to Jev, a decision model he says gets about 70–80% of carefully prompted decisions right [10]. He is “cautiously bullish” on TypeSafe at \$7.5B with about 20 people, even though Cloudflare and Amazon could build a similar product [10].

His exit view is gloomier. \$500M in ARR used to be enough for a safe IPO; now it may take \$1.5–2B. He cites Rory O’Driscoll’s estimate that big tech can buy only about 10 midsize software companies a year. His advice to companies at \$100M ARR: assume there will be no IPO and build toward a strategic buyer [10]. His summary of the week: “42 decacorns in ten months, an \$870M Series A 24 days after launch, and midsize software companies that can’t find buyers” [10].

On cap tables, the Valar/Day One dispute has reached a Delaware filing. It started when one existing investor got an allocation as a “strategic” while everyone else was asked to waive pro-rata rights [10]. Jason’s point for seed investors: pro-rata rights were part of the price they paid at seed [10].

Anthropic dedicates about 20% of its equity to the public good

Jack Clark says that with this announcement plus earlier commitments, about 20% of Anthropic’s equity will go to the public good, and that share can grow. The money will go toward sharing AI’s benefits more widely and funding work

on its risks [11]. Sacks replied: “You’ve put 20% of your equity in an LLC. Whether it serves the public good remains to be seen” [12]. This matters for anyone modeling Anthropic’s IPO cap table.

Infrastructure: where the bottleneck goes after GPUs

Investing in AI estimates that about 120 GW of data-center capacity is scheduled through August 2028, and that historically only about half of scheduled capacity arrives on time. The causes are transformers, gas turbines, and utility interconnection queues that now take 24–48 months [13]. For venture, the author points to long-dated bets: cooling fluids and heat exchangers, small modular reactors and fusion, and photonics [13]. Because power and cooling are becoming a consensus trade, the author sees the least-priced opportunity in software that keeps GPU fleets busy, through utilization and orchestration [13].

Bindu Reddy adds to last brief’s CoreWeave story. She argues neoclouds signed contracts to sell compute below today’s spot prices while paying a premium for labor, power and debt. In her view they should renegotiate or go out of business [14].

Capability and safety research

- **Quantum circuits.** Anthropic reports that Claude Mythos 5, paired with an automated checker, designed quantum circuits that beat the best published ones, by up to 39x on the paper’s cost metric. None has run on quantum hardware yet. IBM separately reported that AI agents found an error-correction decoder design 204x faster than the reference, using \$299 of inference [15].
- **Interpretability.** Goodfire claims polysemanticity came from “ordinary, mostly fixable issues” rather than superposition. It says most neurons can now be understood in models with up to trillions of parameters [16]. This is the company’s own claim.
- **Recursive self-improvement.** A Gladstone AI report, based on conversations with about six insiders at top labs, says most of them expect AI to start improving itself (recursive self-improvement) within 3–12 months, absent government intervention [17].
- **Productivity inside Anthropic.** Anthropic says that as agents spread internally, it produced more code per person, closed more sales per rep, and processed more contracts. Output also rose for the same employees as they adopted new tools [18].
- **Legal.** Keith Rabois says Spellbook is saving a Fortune 100 company \$21M in outside-counsel fees [19]. That is a claim from an investor talking his book.

Agents getting blocked: VCs say it's temporary

A widely shared post listed the walls agents are hitting: Delta, JPMorgan Chase, Schwab, Yelp's paid licensing, and Amazon blocking Grok [20]. Martin Casado calls this friction a normal part of "the adoption cycle" whenever the consumption layer moves [21]. Garry Tan predicts everyone will use AsideAI with their agents "in a year or so" [22]. Casado also argued "don't worry too much about moats" [23]. He compared gen-AI software to molded plastic replacing metal: quality drops, but lower cost grows the market [24].

A cautionary example on AI unit economics: one image app found a single device had made 14,178 generations across nine free trials, about \$1,900 in costs. The app's trial limits were checked per account and its generations were advertised as unlimited [25].

Team signals

Ben Horowitz says a16z specifically looks for original ideas and the ability to build high-quality relationships quickly, since AI already handles "grind-them-out tasks" [26]. Aaron Levie argues that AI's biggest gains go to people with real domain expertise: "the experts just get more powerful" [27].

Sources

1. X post by @thdxr
2. X post by @paulg
3. X post by @AndrewCurran_
4. X post by @DavidSacks
5. X post by @Suhail
6. X post by @Suhail
7. r/Futurology comment by u/Gari_305
8. r/Futurology comment by u/croninsiglos
9. r/deeplearning post by u/No-Conclusion3720
10. Jason's Takes on This Week's 20VC x SaaS: Pro Rata Fights, the First Real AI Price Drop, and Why \$500M ARR Isn't Safe Anymore
11. X post by @jackclarkSF
12. X post by @DavidSacks
13. Beyond The GPU: Second Order Thinking In AI Infrastructure
14. X post by @bindureddy
15. r/Futurology comment by u/Gari_305
16. r/Futurology post by u/Gari_305
17. r/Futurology comment by u/Gari_305
18. r/Futurology post by u/Gari_305
19. X post by @rabois
20. X post by @trikcode

21. X post by @martin_casado
22. X post by @garrytan
23. X post by @martin_casado
24. X post by @martin_casado
25. r/SaaS post by u/buraste
26. X post by @a16z
27. X post by @a16z