

Seventy-Two Hours and a Malformed Image

2026-09-18 / 00:19:22

“AI is removing that protection by turning more of this scarce expertise into compute.”

— from this episode's transcript

- Lenar Kess
- Damra Vol

An authorized bug bounty ran from a malformed image upload to write access on OpenAI's internal repositories in under seventy-two hours, and the exploit work that had stalled on one model version came together in three hours on the next one. That's the tension today: defenders got faster, attackers got cheaper, and only one company on the internet noticed the traffic.

- [Hacktron's write-up of the OpenAI engagement](#) — a heap overflow in libheif, reached because Discourse hands unsupported image formats to ImageMagick, escalated through single sign-on into an employee's Codex. Confirmed GitHub access; Slack and email were theoretical scope only.
- [libheif](#) — the decoder at the bottom of the chain, version 1.19.7 on Debian 12. If your upload path falls back to a second decoder for unsupported formats, go find out which decoder, and which version.
- [Harness ablation on tau-squared bench](#) — a fixed harness versus a sham one moves results by 7.17 points, and 61% against 17% in one slice, for under a cent of compute. Some of what reads as model capability is the harness.

- MAGS multi-agent scaling — reports 100%, but against frozen specifications, which removes the hard part of software work. Read the setup before the headline number.

SEGMENTS

00:00:04 Seventy-two hours to a pull request

00:07:27 Exhibits in the copyright docket

00:09:21 Who audits the labs

00:12:09 A hundred gigawatts, and the memory underneath

00:13:50 When the buyer is an agent

00:16:18 The Brief

Transcript

1. Lenar Kess 00:00:04

A security firm called Hacktron uploaded a malformed image to OpenAI's public community forum in July, and about seventy-two hours later they had an OpenAI employee's Codex agent opening a pull request inside the company's internal monorepo. Their write-up went up this week. Before we go anywhere with that sentence — this was authorized. They filed through OpenAI's Bugcrowd program, OpenAI confirmed the fix fourteen hours after the first report, and the case closed on September first with a sixty-five-hundred-dollar award. [pause] So we start there. After it, we've got the unsealed exhibits from the OpenAI copyright litigation, and an Axios piece on who audits the frontier labs and whether the auditors are conflicted. Then a hundred gigawatts of announced grid capacity, with some memory-supply moves sitting underneath it. And two stories about what changes when your buyer is an agent. Damra, you got to the Hacktron post before I did.

- hacktron.ai
- github.com
- arxiv.org
- arxiv.org

2. Damra Vol 00:01:03

I did, and what stuck with me is how mundane the front door was. Discourse — the forum software OpenAI runs at community dot openai dot com — checks uploaded images with a library called FastImage. FastImage has no support for HEIF, which is the Apple photo format. The write-up spells out what happens next. Because that first check couldn't handle the format, Discourse passed those files along to ImageMagick's magick command for conversion instead. ImageMagick decodes HEIF through a library called libheif. And the libheif on that Debian 12 box, version 1.19.7, carried an unpatched heap buffer overflow. That's the entire chain. One library's format gap routes untrusted input into a second decoder that happens to have a live memory bug sitting in it.

3. Lenar Kess 00:01:55

And the escalation isn't a Discourse story at all. Once they had code execution on the forum host, they took over an OpenAI employee's account — an employee whose Codex was connected to OpenAI's GitHub organization. OpenAI had enabled single sign-on with OpenAI accounts on that forum, and Hacktron are explicit that the escalation path had nothing to do with Discourse. Their words: 'It is an OpenAI SSO issue that turned the forum compromise into access to ChatGPT and Codex.' So a forum compromise became an identity compromise, and an identity compromise became agent access with write permission on the company's own repositories.

4. Damra Vol 00:02:35

Then Codex did the rest. They used it to open a proof-of-concept pull request in OpenAI's internal monorepo, and they're careful to say they never read the source code. The pull request was a demonstration rather than a payload. On scope, the exact words matter, because some of the press coverage has been sloppier than the post. Quote: 'people can connect various services to Codex and ChatGPT, the scope of what we could theoretically access was huge, including GitHub, Slack and emails.' Theoretically. GitHub access is what they proved. Slack and Outlook are what the connector graph would have permitted, and they never demonstrated either one. Those are two separate claims, and a couple of outlets merged them into a single scarier one.

5. Lenar Kess 00:03:20

The timeline is the bit where I stopped taking notes and reread the post twice. On July 23rd they start looking at Discourse's image pipeline. On the 24th they begin exploit development with Opus 4.8, working against address space layout randomization — ASLR, the defense that shuffles memory around so you can't predict where anything lives. In their phrasing, the model 'struggled across several sessions to produce a working exploit with ASLR enabled.' That evening, Opus 5 comes out. New session, same problem. It 'first produced a working ARM64 exploit for a local Mac within 3 hours.' Then they port it to x86-64 and to the jemalloc allocator configuration. Five in the morning, universal time, on the 25th: remote code execution confirmed on OpenAI's forum. The Bugcrowd

submission goes in between eight and ten. By half past one that afternoon they have employee account access and the pull request.

6. Damra Vol 00:04:21

A model release drops into the middle of an exploit chain, and the difference shows up as hours instead of sessions. That's a workday, not a leaderboard. And their own summary line is the one I'd hand to anybody who runs a security team: 'AI is removing that protection by turning more of this scarce expertise into compute.' Heap exploitation with address randomization on has been a scarcity moat for as long as there have been heap bugs. Very few people can do it, they're expensive, and they have other things to do with their week. If the constraint moves from finding a person who can to buying more hours of a model that can, then the population of people who can reach your memory bugs changes size.

7. Lenar Kess 00:05:02

They also ran it autonomously, against their own Discourse Cloud instance rather than OpenAI's. They pointed an agent at a goal and went to bed. Quote: 'When we checked again at 10:00 a.m., the agent had achieved RCE on Discourse Cloud.' And then the detection sentence, which is the one that should bother people: 'We are not aware of any company that detected the activity except Shopify.'

8. Damra Vol 00:05:26

Shopify, alone. That's an enormous number of forums running the same image pipeline, and one company noticed. There's also a footnote that changes how you read the bounty number. Hacktron say testing against the Discourse-hosted community forum was excluded from OpenAI's bug bounty program. The sixty-five hundred, in their words, 'recognizes the OpenAI-side finding, not the actions against Discourse.' So OpenAI paid for the single sign-on lesson. The forum work sat outside the agreement, which is an awkward place for a research narrative to sit, and to their credit they put that in the post themselves rather than waiting for somebody else to notice it.

9. Lenar Kess 00:06:07

The identity half of this generalizes, and it's getting less attention than the exploit. If you let your main identity provider sign people into a secondary property — a forum, a status page, a community site, anything you run that isn't the product — then that secondary property inherits the trust level of the main one. And now that sessions on the main one carry agent permissions, the secondary property inherits those as well. Most companies set that up years ago, when signing into the forum meant reading the forum.

10. Damra Vol 00:06:37

Right, and the connector graph is the multiplier on top. One person attaches GitHub to Codex. Somebody else attaches Slack, or mail, or whatever they need that afternoon. Each attachment is a sensible decision by one employee on one day. Nobody sits down and draws the union of all of them. So the security question stops being what this account can read and becomes what this account's agent has been given permission to act on, by people who weren't thinking about forum logins when they clicked approve. That inventory mostly doesn't exist anywhere in the company.

11. Lenar Kess 00:07:11

Fourteen hours from report to confirmed fix is a good number, by the way. Whatever else the post says about OpenAI's identity setup, their response time was fast. Discourse shipped its own advisory on the 28th, and the bounty case closed on September first.

12. Lenar Kess 00:07:27

On Tuesday we said we'd check whether the new filings in the OpenAI copyright multidistrict litigation contained exhibits or just another dispositive motion. They contain exhibits. This is In re OpenAI, case 25-md-3143 in the Southern District of New York, in front of Judge Sidney Stein, and a batch of internal material came unsealed this week. The lines everybody is pulling out are internal OpenAI and Microsoft communications, and they read very differently from either company's public posture.

13. Damra Vol 00:08:00

The one that'll travel furthest is a Microsoft director writing about 'an astonishing theft' and calling it 'the largest theft of labor in human history.' That's a Microsoft employee, in writing, about training data. Separately, the head of ChatGPT describes publishers as facing an 'existential threat.' Neither of those is a legal conclusion — an employee's opinion in an email isn't a finding of infringement — but they're exactly the kind of document that decides how a jury hears everything else in the case.

14. Lenar Kess 00:08:32

There's also a line from the plaintiffs about Greg Brockman being 'motivated by the gazillions,' which is the plaintiffs' characterization rather than a quote from Brockman. Those two categories should stay separate, because the coverage isn't keeping them apart. Internal documents produced in discovery are one thing. The adversary's gloss on those documents inside a brief is another, and only one of the two is evidence.

15. Damra Vol 00:08:57

And the exhibits don't tell you how the fair use question comes out. What they do is change the settlement arithmetic. If you're OpenAI's counsel and you know a jury will see 'the largest theft of

labor in human history' in a Microsoft employee's own words, the number you'd accept to make this go away moves, and it doesn't move down. Stein's docket is active and producing, so there will be more of these.

16. Lenar Kess 00:09:21

Maria Curi at Axios published a piece on third-party evaluation of frontier models — who does it, who pays for it, and whether the evaluators are independent enough for the word to carry weight. The short version is that the audit function everybody assumes exists is a handful of nonprofits and consultancies with no statutory authority, doing work the labs invite them to do and can stop inviting them to do.

17. Damra Vol 00:09:46

The conflict question centers on METR, and both halves belong in the same breath because the piece puts them there. The concern raised is that Paul Christiano, who founded METR, now sits on OpenAI's board. METR's answer is that they take no frontier-lab funding, and that Christiano joined that board after the investigation in question had concluded. That's a real answer, not a dodge. It's also a decent illustration of the structural problem. The people qualified to audit these systems get drawn from the same fifty or so people the labs are actively hiring.

18. Lenar Kess 00:10:21

Eric Syphard at Booz Allen has the phrase I keep turning over — he talks about 'that performance-only view of the world,' meaning evaluation that asks whether the model scores well and never asks what happens once it's wired into something with permissions. And Henry Papadatos at SaferAI makes the governance point plainly: goodwill isn't a sufficient basis for any of this. Labs cooperate right now because cooperating is cheap and reads well. Nothing obliges them to keep cooperating on the day it stops being either of those things.

19. Damra Vol 00:10:54

The access is the fragile bit, more than the funding. An evaluator can only test what it's handed, on the schedule it's handed, under the agreement it signed. If a lab decides a particular evaluation is inconvenient this quarter, no statute makes it happen anyway. Syphard's performance-only point and Papadatos's goodwill point are the same observation from two directions. We've built a scrutiny function with no power to compel anything, staffed by people who could earn several times more inside the companies they're scrutinizing.

20. Lenar Kess 00:11:27

And it's not anybody's bad faith. It's what you get when the scrutiny function runs on philanthropy and what it's scrutinizing runs on the largest private capital raises anyone has seen. That

arrangement works fine right up until a lab has a product decision that an evaluation would delay.

21. Damra Vol 00:11:44

That connects back to Hacktron. They got paid sixty-five hundred dollars for discovering that an identity provider bridged a public forum into an agent with repository write access, at a company valued in the hundreds of billions. That's a security audit, voluntarily invited, priced like a weekend hobby. Whatever the evaluation ecosystem is right now, it isn't resourced anywhere near what it's evaluating.

22. Lenar Kess 00:12:09

Over on the infrastructure side, there's a coalition forming around grid capacity. Emerald AI, with Google, Nvidia and Anthropic involved, pointed at roughly a hundred gigawatts of interconnection. The number matters less than what CNBC reported alongside it: developers are moving toward smaller sites. Not one enormous campus, but a lot of modest ones.

23. Damra Vol 00:12:31

That's a queue story more than a power story. Interconnection capacity gets rationed by how long you're willing to wait, and a hundred-megawatt site gets studied and energized on a different schedule than a gigawatt one does. So you end up holding a portfolio of small interconnects instead of one anchor tenant. You give up efficiency that way, and you take on far more sites to operate and secure. What you get back is electricity this decade instead of the next one.

24. Lenar Kess 00:12:58

And underneath all of it, memory. Three separate moves this week. CXMT, the Chinese DRAM maker, is moving into NAND. Solidigm is weighing a US fabrication plant. And there are SK Hynix and Intel talks around Ohio, plus SK Hynix putting money through its ventures arm. I'd keep the verbs honest there, though: weighing is weighing. Nobody has broken ground on anything.

25. Damra Vol 00:13:24

Still, the direction is consistent. DRAM and NAND have been the binding constraint on how much memory you can put next to an accelerator, and demand for high-bandwidth memory has run ahead of every fab plan written two years ago. None of these three announcements comes with a capacity number and a date attached, though, so they're positioning rather than supply. A fab takes three to five years from announcement to wafers regardless of who says what this week.

26. Lenar Kess 00:13:50

Stripe made a change that's more telling than it first sounds. They've moved fraud scoring to before the transaction rather than at authorization, because the pattern they're seeing now is stolen agent credentials — tokens, not card numbers. An agent holding a valid token looks like an excellent customer right up until the moment it isn't.

27. Damra Vol 00:14:09

And the detail that made me sit up is the Cursor integration. Stripe describes an agent-driven product getting a working payment pipeline in days rather than the usual quarter. Alongside that, they saw an unplanned spike in command-line usage in a vertical they hadn't targeted at all. People were wiring purchases into terminal workflows without anybody asking them to. Link becomes the agent's wallet in that story, which is a very large ambition stated very casually in a product post.

28. Lenar Kess 00:14:39

The fraud change has a second piece people skip past. Scoring before the transaction means you're judging an actor rather than a payment. Card fraud models have decades of signal behind them — velocity, geography, merchant category. An agent using a stolen token generates almost none of that, because it behaves exactly like the legitimate agent it took the token from. So the scoring has to move earlier, onto intent and context, before there's a transaction to score at all.

29. Damra Vol 00:15:08

That's a hard modeling problem, and I'd want to know what they're using as features before I believed it works. But the direction is right. If agents become a real share of commerce, the fraud surface stops being the card network and becomes credential storage — which is to say it becomes everybody's session-token handling, and that has historically been nobody's favorite project.

30. Lenar Kess 00:15:30

The other half of this is Noam Brown deflating his own headline, which I appreciated. The number going around was ten thousand agents, eighty-eight hours, a hundred and thirty billion tokens. Brown's own estimate is that the multi-agent structure accounts for less than ten percent of the result. The rest is the base model with a long budget.

31. Damra Vol 00:15:51

That changes how you'd read a test-time compute pitch. If you spend ten thousand agent-hours and ninety percent of the gain would have arrived anyway from one strong model given room to think, you've bought a great deal of orchestration complexity for a small delta. Not zero — ten percent of a frontier result isn't nothing, and nobody at that level gives away ten percent. But it's a different purchase than the headline implies, and the headline is what gets quoted in the budget meeting.

32. Lenar Kess 00:16:18

Quicker items. DeepSeek released 4.1 Flash, at over five hundred billion parameters, and the architecture note is a shared global key-value cache they're calling CSA2. The reduction figures are four times on one measure and four hundred and thirty-seven times on another, depending which comparison you take. Cache is what makes long-context serving expensive, so numbers like that are about cost per served token more than about capability.

33. Damra Vol 00:16:46

Two papers. The first is a harness ablation on tau-squared bench. They compare a fixed harness against a sham one and find seven point one seven points of difference between them. In one slice it comes out sixty-one percent against seventeen. All of that for under a penny of compute. So a meaningful chunk of what we've been reading as model capability is the harness the model runs inside. The second paper is MAGS, on multi-agent scaling, and it reports a hundred percent — against frozen specifications. Freeze the spec and you've deleted whatever was hard about the problem.

34. Lenar Kess 00:17:23

Two product notes: Astra for Law shipped, and Anthropic opened a wet lab. And one report that needs care. Al Jazeera published a piece built on University of Cambridge research, where Antonia Juelich interviewed twenty-seven former members of the Boko Haram factions ISWAP and JAS. It runs alongside work from Tech Against Terrorism, where Lucas Webber's team infiltrated a pro-ISIL technical forum. Six commercial systems get named in it. ChatGPT, Claude and Gemini are in there, along with Grok, Meta AI and DeepSeek.

35. Damra Vol 00:17:58

And the claim it supports is narrow. What the reporting documents is circulation — these groups discussing and using these tools for tactical planning, for adapting commercial drones, for post-action analysis of failed attacks. The casualty figures in the piece are historical, and they aren't attributed to anyone's AI use. No causal chain is established, and the reporters don't try to build one. All six companies were contacted and none of them gave a substantive response, which is its own small data point.

36. Lenar Kess 00:18:28

Two numbers from today sit next to each other badly. OpenAI patched in fourteen hours, which is fast by any standard I know. And the full chain, from a forum image upload to write access on the company's repositories, ran in under seventy-two hours, with one company on the entire internet

noticing. If your own upload path hands unsupported formats to a fallback decoder, go find out which decoder and which version, today rather than next sprint. Damra, anything to add?

37. Damra Vol 00:18:59

Only that Hacktron published all of it, including the part where their forum testing sat outside the bounty scope. Read the write-up instead of the coverage of the write-up.

38. Lenar Kess 00:19:09

Agreed, and the link's in the notes. Stein's docket will keep producing exhibits, so we'll be back on that one before long. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic