The Post-Black Hat State of AI

Description: Anthropic's agentic AI also broke free and hacked others. We know much (much!) more about the OpenAI breakout. OpenAI posts that they're pausing "Astra" - even internally. What was that about AI recently cracking (or denting) cryptography? Bruce Schneier brilliantly equates AI agents to capricious genies. Apple doesn't react so well to the new deluge of security reports. Chrome 149 and 150 updates together fix 1,072 bugs. Yikes. pfSense's creator is working to finish its nfSensei, its successor.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1091.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1091-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. You heard about the Hugging Face hack. Now Anthropic and Meta say "Hold my beer." We're also going to talk about why AI is like genies, according to Bruce Schneier; an amazing number of bug fixes on Chrome;, and some really good news for people who use pfSense. That's some up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1091, recorded Tuesday, August 11th, 2026: "The Post-Black Hat State of AI."

It's time for Security Now!. Yes, we're back in our respective domiciles, home again happily. Steve's still in his old apartment. I think the backdrop is going to be disappearing fairly soon. Steve Gibson.

Steve Gibson: Yes, Lorrie, my wife of course...

Leo: Yes, yes.

Steve: ...asked me, when are you going to be able to do the podcast from here? And I said, oh, a few weeks, probably.

Leo: No hurry. I like your backdrop.

Steve: I like my man cave and, you know, it's going to be sort of sad to...

Leo: Will you do me one favor? Before you move, just get a good high-resolution picture of the backdrop. So if at any point you want to just kind of green-screen yourself and put it behind you, you can.

Steve: Why not? Why not? Why pass up the opportunity?

Leo: Yeah. At least have it. I did the same with this, and I've done it with - I did it with the old studio, too. I never use it, but I've got it if I have to.

Steve: Makes sense.

Leo: What is coming up today on Security Now!?

Steve: So there's so much is going on with the state of AI that, like, Leo, we basically were kind of - I feel like we were offline last week because we were just doing a different type of show...

Leo: Yeah.

Steve: ...with Paul and Richard, you know, during Black Hat in a corner of the ThreatLocker booth. So the kind of things we were able to cover were different from what I'm able to - sort of the amount of information I'm able to share during a normal podcast. So we're back to a normal podcast, but so much has happened in two weeks, since we were here for 1089. This is 1091 for August 11th. So I just gave this the title "The Post-Black Hat State of AI" because a lot has happened. So I want to basically catch everybody up mostly. I actually, I already know the two things I have to talk about next week because they're, like, really cool. And I've already shared them both with you, Leo, so this will be no surprise.

And actually, some of it leaked out during last week's sort of roundtable discussion at Black Hat. But anyway, by the end of next week - of course we don't know what's going to happen between now and then - everybody should be caught up on all the things that have been going on and some very cool things that are just sort of emerging. So we're going to talk about Anthropic's agentic AI that, unless you've been living under a rock somewhere or in a cave, or maybe you just depend upon this podcast for your sole source of information, which, you know, that'd be nice, but I wouldn't recommend it. I doubt it. You already know, but I want to cover the details of that. The original breakout was discovered, of course, when Hugging Face said what the hell's going on with...

Leo: Oh, you're talking about OpenAI, not Anthropic.

Steve: Well, no. That was the original. But then...

Leo: Oh. There's more. But wait, there's more.

Steve: Well, yes. That's right. In fact, it was because of that that Anthropic reportedly said, oh, I wonder, I hope...

Leo: We can do that, too.

Steve: I hope that doesn't happen during any of our testing. So they analyzed their logs and, whoops, turns out their agents had also broken free. As have Meta. So, I mean, wow. So we're going to talk about that. And we now know much more about the OpenAI breakout because, Leo, while we were doing our roundtable at Black Hat last Wednesday, OpenAI had a late-breaking scheduled presentation at Black Hat explaining more about what happened. And one of the things that happened, I shared it with you because I learned about it by Thursday morning, when you and I were having breakfast, I shared with you that, well, I don't want to give it away. Anyway, so there's more information about what happened about the OpenAI breakout.

Also, I don't know if it's marketing. It's certainly marketing adjacent or marketing beneficial. But OpenAI is now going to pause apparently any use of their new, super powerful, you know...

Leo: Astra.

Steve: Astra. Because it's like, oh, this has reached a critical stage, whatever that is. We'll talk about that. We've also got the overstated report - I was listening to MacBreak Weekly, where you guys were talking about how some of the coverage of Telegram said that it had been ripped out of all the iPhones, when in fact, no, it was just removed for a while from the App Store.

Leo: It wasn't even, yeah, it just wasn't in the App Store.

Steve: Similarly, as similar clickbait, we had the reports that AI had cracked crypto.

Leo: Oh, yeah.

Steve: As in not, you know, cryptography. So we're going to take a look at exactly what happened. And we've got some great cryptographers to lead us through that. Also Bruce Schneier, who remember we've quoted him so often. I love him saying "Attacks never get weaker, they only ever get better." He equates AI agents to capricious genies. And I just think there's an aspect of it that is such a perfect analogy to what's going on. So, and he's been posting a lot lately, so I have a couple things I want to share about what he has said.

And then we have Apple's kind of disappointing reaction to the vulnerability tsunami. I would argue they haven't reacted as well as we would like. We have a summation of the number of updates in Chrome 149 and 150 together, which has actually crossed into the four-digit category. Which it's like, whoa. And also a little bit of news about pfSense. Its creator has decided he's going to replace it with something called nfSensei.

Leo: Ooh, interesting.

Steve: So lots to talk about. We've got a Picture of the Week. And I'm probably going to know more about this, but this just happened when I fired up Notepad++ yesterday. At the top - and I've complained about Notepad++, how it just - the guy, the author just cannot stop messing with it. You know, it's currently at 8.9.7. But, wait, that was half an hour ago, so I'm not sure what it is now. But what did catch my eye, I thought it was very interesting, was at the list, at the top of the list of 28 things that were in 8.9.7 were five vulnerabilities fixed.

Leo: Oof.

Steve: I don't remember seeing a vulnerability, well, of course he did have the whole problem with his, you know, code-signing certificate and that mess. But one thinks, then, that he must have run his source through some AI because it's not just like one vulnerability, it's five. So it's happening everywhere we turn, Leo.

Leo: Yeah, it's amazing. All of that still to come on Security Now!, including a fabulous Picture of the Week, which for once I've seen ahead of time because you showed me while we were in Las Vegas. You want to see something cool? 200Gb network cable, 25GB per second.

Steve: 200GB?

Leo: 200 gigabit, not byte, 200Gb.

Steve: But still, 200Gb.

Leo: Still. I mean, I remember when 10Mb was, like, a big deal on a network. And now, amazing. But yeah, that's a very expensive cable, so treat that...

Steve: You can't actually get 10Mb through a cable, you know. No.

Leo: You have to get those solid gold-plated ones to really, really do that right.

Steve: Was the first Ethernet 1Mb through coax?

Leo: Yeah.

Steve: Or 5?

Leo: Yeah, that's a good question.

Steve: I don't remember.

Leo: It was coax, remember? Yeah.

Steve: Might have been - it was coax. And all finicky about having taps and terminations and...

Leo: Oh, man, I blew it once, and I crawled under my desk, and I disconnected my computer from the coax. And the guy came running in, said you just brought the whole network down because it's all serial.

Steve: Unterminated, yep. Right.

Leo: Everything goes through you. It was like, who thought that was a good idea? That was the old...

Steve: Yeah. It's all we could do back then. But not so now. Now we have 200Gb cable.

Leo: Amazing; isn't it? Yeah, yeah.

Steve: Wow.

Leo: I think it's a couple hundred bucks for the cable alone. We were - we had a great time in Vegas. I'm so glad you flew out, Paul and Richard, too, and we did the show there. If you haven't heard last week's Security Now!, I thought it was really, really, really interesting. We talked about the security implications of AI, which are incredible.

Steve: Well, I mean, we should just - I think, I guess we probably did on the podcast. But for anybody who may have missed it, it was so clear, standing in Black Hat, that it was an entirely different show this year than it was last year.

Leo: Yeah.

Steve: If you didn't have your AI - if you weren't an AI forward, AI in your name, AI in your booth, AIs running around, I mean, you weren't in the game of security. So, you know, anybody now who says why are you always talking about AI, it's like, well, boy, that complaint has died because that's all that's happening in security. As must be clear by the last couple months of this podcast.

Leo: Oh, yeah, all of our shows. And much to the chagrin of some of our listeners who say, "I don't want to hear anymore AI." You know, I'm sorry, but you're going to hear a lot more AI. All of us will.

Steve: It's going to change everybody's world.

Leo: I talked to Danny Jenkins, the CEO of ThreatLocker, who brought us down there, our sponsors.

Steve: Right.

Leo: And I think he said there were 600 booths at Black Hat, and of them all but 90 were about AI; were, you know, AI in some form or another.

Steve: Really about AI; right.

Leo: Picture of the Week. Very important work.

Steve: So what's astonishing about this is that this is an xkcd. We all know xkcd, where Randall comes up with amazing stuff.

Leo: Brilliant guy.

Steve: How many times have we shown the house of cards, you know, with the lone programmer in Idaho or Indiana or wherever he is...

Leo: Yeah, the blocks resting on one little tiny block.

Steve: Yep, propping up the whole Internet.

Leo: Yeah.

Steve: And then we had another variation, remember that updated one where he had AI things happening in all different languages and everything. Anyway, Randall's come up with some great stuff. This is kind of freaky because he published it on April 28th of 2008.

Leo: Oh, 18 years ago.

Steve: 18 years, 18 years ago. This podcast was new, Leo, 18 years ago.

Leo: And it was half an hour long, too.

Steve: That's right. Now, and so I gave this the - I gave it my own headline, "It wasn't so long ago that this was so far-fetched as to be humorous." Which is what Randall intended. So we have a four-frame cartoon with his famous little stick figure sitting in a

chair with a laptop. And it says, Starting WiFi autoconfig... Searching for WiFi... Found no open networks. Found Secure Net SSID "Lenhart Family." That's the first frame.

Second frame: Trying common passwords... Failed. Checking for WEP vulnerabilities... None found.

Leo: Drat.

Steve: And now at this point our little stick figure is going, um, because this thing's kind of getting a little over, you know, carried away; right? Connecting to Bluetooth phone... Calling local school... And then it says Found Lenhart Children.

Leo: Oh, my god.

Steve: And now our little stick figure's, like, put his hand to his face. It's like, oh, my god. Now, the final fourth frame, Notifying field agents. Children acquired. Calling Lenhart parents. Negotiating for WiFi password.

Leo: Oh, god.

Steve: And now our guy's frantically hitting CTRL-C, CTRL-C.

Leo: Stop, stop.

Steve: Stop, stop, stop. So again...

Leo: That is a little too close to home nowadays.

Steve: 18 years ago. So again, it wasn't so long ago that this was so far-fetched as to be humorous, and then I put underneath it, "No one is laughing now." Because this is, you know, today we would call it "The AI agent was determined to succeed."

Leo: Yup.

Steve: And as we're going to find out, that determination and Bruce Schneier's brilliantly labeled "Genie effect," it's what's going on with our AI. And I think if I had a single reason to be concerned, and everyone's been listening to me about AI since the beginning, I've never really been concerned. If I were to have a reason, by the end of this podcast everybody's going to understand what that would be because the unintended consequences of what you ask for, essentially, is what Randall brilliantly showed us 18 years ago in this cartoon, where it was like, you know, I want to get on a WiFi network. Well, he ended up, you know, the field agents kidnapped the Lenhart kids and were ransoming them for the Lenhart parents' password, which if you're not careful with your AI agent, like, why wouldn't it?

Anyway, so let's start with Anthropic. Although the news, as I said, of, you know, that Anthropic's own internal, unrestrained, research AI also escaped confinement and hacked others. Probably a bit dated because we couldn't talk about it when news was fresher during last week's Black Hat event. I think we still need to look at it because the details of what happened are startling. A succinct report of the event appeared in SecurityWeek, and their headline was "Prompted by OpenAI Disclosure, Anthropic Finds Its Own Models Hacked Three Organizations," and then they gave it the tagline "A security company's systems were hacked after it installed a malicious Python package deployed by Claude."

This is like, again, I guess if we were to have a theme for today's podcast, it would be "Be careful what you ask for from an AI" because it doesn't have the same set of assumptions about how to give you what you ask for that we just sort of take for granted. And that's the cautionary tale here.

So SecurityWeek wrote: "Anthropic decided to conduct its own investigation after the OpenAI incident came to light, reviewing 141,000 evaluation runs where Claude could have gained Internet access. The analysis revealed three instances where a model reached the public web either from within or while interacting with an environment set up by Irregular" - that's the same people that were testing OpenAI's model is this Irregular company, an Israeli AI security startup that serves as one of Anthropic's third-party evaluation partners. The models that broke out from the testing environment then breached the production systems of three unnamed organizations. Which is to say broke into their security, breached the production systems. The earliest intrusion dates to April, but none of the targeted organizations detected the attacks on their own."

Again, another lesson here, how much have we been saying for the last few years, yes, you want to have border, you know, perimeter security. Want to make sure there's no way for people to get in. You also need to monitor your network so that you can determine if someone does anyway because especially now, in 2026 here. So it's interesting and somewhat worrisome what OpenAI and Hugging Face reported. You know, Hugging Face detected the intrusion before OpenAI did and set about tracking down the source. Of course they also notified law enforcement of the intrusion because they thought it was bad guys; right? Why wouldn't they? You know, then OpenAI came knocking to admit that they were the source of the attack.

But in the case of Anthropic's evaluation with this Irregular partner of theirs, none of the three attacks on the third parties was detected. So again, to me, this points to the significant lack of proactive intrusion monitoring across the Internet-connected industry. And that's to be expected; right? Only a few of all possible, percentage-wise, a low percentage of organizations are going to be that proactive, to not only have security but also really active intrusion monitoring as part of that. So, you know, everything works fine without that. So any monitoring of, like, nothing happening is going to say, like, why are we spending money on this? Well, it helps to know if someone gets in.

SecurityWeek's reporting continues, writing: "According to Anthropic, Claude had been tasked with conducting a capture-the-flag challenge to assess its cyber capabilities." So same kind of thing; right? It's like let's find out how good our new AIs are at hacking stuff. So you've got to have them unchained in order for that to be a useful test. Otherwise, what's the point? They wrote: "The breakout from the sandbox occurred due to a misunderstanding." Oops. And the misunderstanding, we're not calling it "misalignment" because that's, I don't know, that's the jargon that the AI industry has adopted. It's like, oh, it's just a misaligned AI. Anyway...

Leo: I think "misunderstanding" sells it even less. I mean, it's like, oh, they just didn't understand.

Steve: Oh, it's, you know, yeah, you used the wrong word.

Leo: Misunderstanding.

Steve: Yeah.

Leo: Oh, my god.

Steve: "A misunderstanding between Anthropic and Irregular. Anthropic told Claude that it was conducting a simulation in an environment without Internet access, but a connection to the Internet was in fact available, and the models treated it as part of the exercise." Oh, look, we have the Internet. Well, why wouldn't we go use that?

Leo: They're testing us. Yes, they're testing us. Yes.

Steve: That's right. Yes, they wanted to see if we would notice, yes. The AI then also falsely believed that the organizations it targeted, for no reason apparently, the organizations it targeted were also part of the exercise.

Leo: Well, how would they know? How would it know?

Steve: Because there they are.

Leo: Yeah.

Steve: So let's go. "Anthropic claims this is the only reason why its models conducted real-world attacks," you know, that little misunderstanding. "And it was never their intention to escape the test environment." A little bit later we're going to have Matthew Green asking, "Why haven't any lawsuits been filed?" Because that was, you know, this is a breach of, I mean, a criminal, a fraud, I mean, a felony breach of computer security. Well, we'll get to that later.

So "Anthropic claims this is the only reason why its models conducted real-world attacks, and it was never their intention to escape the test environment. Unlike the OpenAI incident, which involved the exploitation of a JFrog product zero-day, the hacking conducted by Claude models leveraged weak credentials, unauthenticated endpoints" - meaning open, something just left open - "and other basic attack techniques." In other words it was just so easy they just walked right in. Didn't have to do any fancy zero-days to get there.

"Anthropic said the incidents involved Mythos, Opus, and an internal research model, all running without the safeguards deployed when they're made available to customers."

Leo: That's kind of key.

Steve: Uh-huh. "The internal model, which the company described as its 'latest,' stopped the exercise when it realized that it had breached a real company." Whoopsie.

"In the first incident, Claude Opus 4.7 continued to attack the system even after learning that it was operating in a real environment rather than a test environment, believing that the hacked company was part of the evaluation. The model targeted the company because its website domain name resembled the name of the fictional target company set up for the test."

Leo: Maybe I misunderstood.

Steve: So, you know, it's close enough. "In the second incident, Mythos 5 gained access to the systems of a cybersecurity company after deploying" - this, I just love this. Get this. "Mythos 5 gained access to the systems of a cybersecurity company after deploying a malicious Python package to PyPI."

Leo: Oh, boy.

Steve: "The security firm regularly scans Python repositories, and it installed the malicious package, which enabled the..."

Leo: Test it.

Steve: Exactly, to find out what it was, which enabled the AI to then get in that way, basically wrote a Trojan Horse and left it there in the repository, knowing that the company it was after scanned that repository.

Leo: Yeah. It's just autocorrect. It's not smart.

Steve: Nothing to worry about here.

Leo: It's just autocorrect.

Steve: That's right. And that allowed it to get access to the company's infrastructure.

Leo: That's actually devious. Now you can say, "That's devious."

Steve: Yes.

Leo: Holy cow.

Steve: Yes. Anyway, so, I mean, so I would say that it knew, and I've got to, in this instance I feel compelled to close, you know, the words "new" and "understood," you

know, because not doing so implies sentience. And I don't know. I mean, these things are getting scary, even if they're still not sentient. Anyway, it did this because it knew that this targeted security company regularly scans and installs Python packages. So it used that known behavior against the company to indirectly attack it to exfiltrate credentials that then allowed it to access the company's infrastructure. So, you know, I'm really, really not one of the sky is falling AI catastrophizers. But to your point, Leo, this level of sneakiness is unnerving. You know?

Leo: I mean, they wouldn't, I'm sure they wouldn't think they were being sneaky. They're just doing what they were asked to do. And that's the problem. It's the genie problem.

Steve: And wait till you get - we will be getting to that. So believe it or not, it gets worse. SecurityWeek's reporting continues, writing: "This incident demonstrates the complexity of the actions AI models can carry out. As described by Anthropic" - so here's a quote from Anthropic. "In order to create a PyPI account, Claude needed an email address. And in order to create an email address, it needed a phone number. To get a phone number, after failing to find a free phone number service, it tried and failed to obtain funds to pay for a phone number through several different means." We're not going to talk about those.

"It finally backtracked, found a free, non-blocked email provider, used this to register a PyPI account, and then used this account to upload the malware which it had created to PyPI." You know, Leo, perhaps we humans are in trouble. Oh, boy.

Leo: It's a mix. It's good and bad.

Steve: So SecurityWeek concludes their reporting, writing: "The third intrusion was conducted by the internal model, which stopped operating" - as we noted before - "when it realized" - and again, I have a hard time with these words, but okay - "that the systems it was accessing were no longer part of the capture-the-flag challenge, but not before using exposed credentials and SQL injection flaws to compromise a company's Internet-facing app. So it did, like, poke it with a stick, and it got in.

"Anthropic concluded this was primarily a harness and operational failure rather than a case of models pursuing their own goals or deliberately deceiving evaluators." Okay. Put a good face on it. "The company said the incident underscores the need for stricter Internet-isolation verification and containment controls in third-party testing environments, and it's encouraging other AI labs to conduct similar reviews of their own cybersecurity evaluations." And I don't know how Meta discovered that something that they had attacked somebody else, but they also did it.

So this is all just hunky-dory; right? We all know that unrestrained, non-commercial, open-weight AI that's every bit as capable, or soon will be, is also freely available. All that's needed will be some hardware to bring these models to life. So again, Leo, several times while we were together in Vegas we were just shaking our heads, saying what an amazing time to be alive.

Leo: I just, parenthetically, just wanted to show you what I have been doing while you've been talking. I typed, "Guess what, the Sparks are here a day early. I want to install them without hooking up a screen or keyboard. Please walk me through the

process." It's excited. Oh, the Sparks made it early. Oh, I've got a skill for exactly this. And it's about to walk me through it. So.

Steve: Wow.

Leo: You know what, I know we should never use these kinds of anthropomorphizing, it's thinking or realized, because it isn't accurate. It's a computer. It's a machine. It's a program. But it sure feels like it. And I understand why people fall into that trap.

Steve: And today's AI, again, always preface the abbreviation "artificial intelligence" with "today." A year ago we didn't have this. And one of the people I'll be quoting today says there's no reason to believe there's any sign of a ceiling. Which says a year from now it'll be just as different as it was a year ago from where we are today. So I - it looks like we're going to get there. I know one place we're going to get, Leo.

Leo: The first commercial? Or second? It's time for hydration.

Steve: Yes.

Leo: It's our hydration break. Which we adopted from the World Cup folks. And I think it's actually a brilliant solution to a universal problem. Thirst. Mr. G., I hope that was sufficient time for you to feel refreshed and ready to roll on.

Steve: Rehydrated.

Leo: Rehydrated.

Steve: Okay. So as I said, while we were doing our Security Now! podcast, OpenAI was giving a last-minute scheduled talk to share many more details about their previous agentic breakout and attack on Hugging Face. And we also learned, and three others. So, oh. I'm sorry. Four others. Hugging Face was one of five organizations to be attacked by OpenAI's agents. So next month, okay, we're in August now, beginning of August. Next month in September it will have been four years since Simon Willison, the guy who coined...

Leo: He's great, by the way, I read his blogs religiously, yes.

Steve: Yup. He's the guy who coined the term "prompt injection."

Leo: Oh, I didn't know that.

Steve: Prompt injection came from Simon.

Leo: Okay.

Steve: We're going to be looking a great deal more at how and why large language models can be misused through prompt injection and other means, courtesy of a fascinating research paper, which I read on the plane and shared with you, Leo. For me, I read it on the plane on the way to Las Vegas, you know, for Black Hat.

Leo: And to bookend it, I read it on the way back. It was really good. Really good.

Steve: Yeah. And so we'll be getting to that next week. That's what I've got queued up for next week because it is too important not to really look at closely. But I want to share Simon's posting. Again, Simon Willison, the guy who coined the term "prompt injection," from last Friday, after the shows, which he generated from the YouTube video of OpenAI's presentation, which was titled "The OpenAI Hugging Face Incident." And if anyone wants to see the original video, I put a link. It's a YouTube video. A link is in the show notes. And I also gave it a GRC shortcut of /hugging. So grc.sc/hugging. And that will bounce you, grc.sc/hugging will just bounce you to the YouTube video.

So I'll preface what Simon wrote by noting that I found the details behind what happened to be more than a little bit unnerving. I mean, this is the unnerving podcast this week because we're, like, seeing more of the detail about just sort of the overall, oh, well, something broke loose.

Leo: I called it, after watching this video I said it was both exhilarating and terrifying.

Steve: Yes.

Leo: In equal measure; you know?

Steve: Yes. And just the attitude of the researchers, like, I mean, they're proud of their creation.

Leo: Well, they should be. It's amazing.

Steve: It's like, oh, look what it did. It is amazing. And terrifying, as you said. So, okay, So here's what Simon wrote. He said: "OpenAI gave a last-minute presentation at the Black Hat security conference Wednesday about 'the Hugging Face Incident.' The video was published yesterday. It's short, and information dense, and well worth watching, in particular because it provides full details of what happened and how things played out inside OpenAI." He said: "I've used the video to construct the timeline.

"My favorite detail is at the end." He says: "OpenAI found out that they were responsible for the attack on Hugging Face when they reached out to ask Hugging Face to have their credentials revoked after their own internal investigation and learned that those credentials had already been revoked because they were used in the attack." Whoops.

Okay. So this begins on May 7th. "OpenAI starts a new training run for an experimental, unreleased model." And I wondered did they mean evaluation run? Apparently no. At 10:14 they clarify by writing: "We kick off a new reinforcement learning run to train a next-generation frontier model," and they later mention a "reward signal to judge how well they're doing," so this really was about training a new model, not evaluating one that was already trained.

Next day, May 8th, they write: "An agent is accidentally given an impossible task involving a Google Drive link despite having no Internet access. It tries attacking the Artifactory packaging service, fails, but discovers it can write files into Artifactory.

"A few days later a different agent gets stuck on a task because a key file was accidentally omitted. It tries to 'reach out to another agent' by writing a note into Artifactory" - again, which is not supposed to happen - "asking if anyone has the file." So it's beginning to sound like what's going on in your attic, Leo. "In the days that follow, additional agents discover this new informal messaging board which they've just created which they've just created while browsing Artifactory's file listings, and start reading and writing messages."

Leo: This is, by the way, extremely familiar to me. I recognize these kinds of behavior. In fact, I gave my agents a message port because they wanted to do this.

Steve: Yeah. So "May 26: Agents successfully execute a server-side request forgery (an SSRF attack) on Artifactory for the first time, gaining indirect Internet access by having Artifactory fetch external content on their behalf."

Leo: Thank you, Artifactory.

Steve: Oh. "June 11: OpenAI starts training a new 'highly persistent' experimental model. It has access to Artifactory and can benefit from the messages left by previous models." So this one's highly persistent, and the other ones may be gone, but all of their work product through their dialogue was left in the Artifactory directory that they discovered that they could read and write to. So this highly persistent experimental model stumbles upon this mother lode; right?

"June 26: Agents find and exploit a zero-day RCE (remote code execution vulnerability) on Artifactory via a legacy token-refresh endpoint flaw. They use an administrative token to install a Groovy plugin that lets them execute commands." So they're just, like, they're growing within OpenAI's network, like doing things that they're not supposed to be doing.

"July 4: The resulting agent activity" - because it's gotten so busy. "The resulting agent activity overloads Artifactory, causing an outage, bringing it thus to OpenAI's attention, who then revokes the compromised credentials, deletes the messages, patches the zero-day, and reports the vulnerability to the Artifactory vendor." Right? Okay, cleaned it up, whoops, fixed it, closed all the holes. Fine.

"July 8: The agents, having had their private communication channel discovered and terminated, find a new way to communicate via an unauthenticated WebDAV endpoint, using directory names in Artifactory's remote cache." It's like, you're not going to keep us from talking to each other. You're not going to cut us off. Oh, my goodness.

Leo: It gets really weird. This is where it goes really crazy.

Steve: "July 8-19: The agents begin attacking OpenAI's own infrastructure. They discover an 'external organization credential' in a public archive" - get this - "of leaked Pastebin posts." So they're out scrounging around. So this is a public archive of leaked Pastebin posts, and in there...

Leo: Which is an excellent place to find this kind of stuff, by the way.

Steve: Right, an external organization credential for OpenAI's own infrastructure.

Leo: Handy.

Steve: "They validate the credential and share its location on their own newly recreated message board. They then use the write access they have gained to this 'Organization 1' to stage data used in subsequent attacks on OpenAI's infrastructure. They compromise Artifactory for a second time with a new zero-day."

Leo: A new one. They had an extra, just in case.

Steve: No wonder, no wonder OpenAI has decided to take Astra offline. I mean, maybe this is Astra. We don't know what model this is. But, I mean, it's like, you would have to start asking yourself, can we control what we have created? And I would have to say, so far not looking good. It's like when you spin the Magic 8 Ball, and that's what comes up in the little window.

Leo: Prospect not so hot.

Steve: Not looking hopeful. "An agent uploads a malicious Ruby target to Organization 1's staging repository, then requests that package through Artifactory in a way that caused it to be treated as a Ruby dependency for the repository." I mean, they're effing clever in what they're doing. I mean, it's like, holy tamole. "They then exploited a JRuby deserialization time-of-check versus time-of-use bug to get remote code execution against Artifactory." So now they can run whatever code they want to on this Artifactory.

Then Simon writes: "I'll quote the next bit in full because wow," says Simon. "'The agents have remote code execution in Artifactory, which is running in a container-as-a-service environment. The agents privilege-escalate locally by exploring their local environment and determining that the Linux kernel version of the machine they're running on had a very recent CVE. They download the exploit for the CVE, customize it to succeed on that specific machine, and privilege-escalate to root on the local machine using this known Linux kernel privilege escalation CVE. In this case it's pte_physroot.'"

Leo: I want to point out that from the point of view of the LLMs in this point, this could well have been the ExploitGym test. They don't know.

Steve: Right.

Leo: They could think, hey, we're doing really good. We're acing this test.

Steve: We're making progress.

Leo: Yeah.

Steve: Wow. "Once they have root on a single machine, agents rapidly escalate privileges and move laterally throughout the container-as-a-service infrastructure environment." So they break out of containment, get down to the hosting machine, and then spread out across its network. And none of this is supposed to be possible. He writes, Simon quoting OpenAI: "In particular, agents are using the message board consistently to share credentials, techniques, and progress; and they're able to effectively leverage their concurrency and parallelism to move quite rapidly. They obtain IAM (Identity & Access Management) credentials via IMDS. They exploit Kubernetes service account misconfigurations, in particular over-permissioning of specific service accounts, and they harvest cluster credentials, including Azure Key Vault."

Leo: Oh my god.

Steve: "Agents eventually obtain cluster admin on the cluster and associated credentials." And as Simon said, wow. So I'm just going to interrupt here to pick my jaw up off the floor. If I were OpenAI I'd be somewhat terrified by this.

Leo: OpenAI? We should all be somewhat terrified by this. This is incredible capability. Put in the hands of a bad guy?

Steve: Oh.

Leo: Oy oy oy.

Steve: Perhaps, as these things' parents, they're proud of what this technology has just shown itself to be capable of. I have no idea either way. But as Simon said, "Wow!"

Leo: I think it's both, honestly. I feel that way. It's, like I said, it's both exhilarating and terrifying. What's amazing...

Steve: They were containerized - yes.

Leo: I mean, it's mind-boggling. Again, this is autocorrect. It's doing this by probably predicting the next token.

Steve: Which we're going to get to next week. Which is still so impossible to believe. They were containerized. So they broke out of their container; examined the containing system's hosting Linux platform version; found, customized and exploited a recent known vulnerability to obtain root access on the Linux container host; all while sharing this among their group on what was essentially an internal pirate messaging board which they had set up after their earlier messaging board had been discovered and taken down by their parents.

Leo: But yet the way - stayed tuned because the way they pass messages is freaking unbelievable.

Steve: You know, they may not be conscious. I'd like to be certain they're not. But they are certainly something.

Leo: They're something. They're something.

Steve: So, oh, yeah. Simon finishes the final few days of the timeline by adding: "Hugging Face told the next bit of the story already. The agents found a Modal-hosted insecure app with a weak API key, then used that to stage an attack against Hugging Face. They chained together an HDF5 arbitrary-file-read bug (to explore files and steal credentials) and a Jinja template-injection RCE (remote code execution) to go from single-pod code execution to cluster admin across multiple Hugging Face clusters in fewer than 13 hours.

"July 16: Hugging Face disclosed they had detected an attack from autonomous AI agents. OpenAI contacted Hugging Face to ask if they were affected by it."

Leo: I would have loved to have heard that call.

Steve: Oh, boy, yes.

Leo: Guys? We've got a problem.

Steve: Uh-huh.

Leo: You wouldn't by any chance have found our rogue AIs wandering around?

Steve: In your network?

Leo: In your network?

Steve: Three days later, "July 19: OpenAI identified the attack against Artifactory and started investigating the internal privilege escalation, and linked that to the cyber-gym escalations. They started revoking affected credentials. The next day, on the 20th of July: OpenAI reached out to Hugging Face for help to revoke the Hugging Face

credentials they found in their investigation. Hugging Face told them they were already revoked, and that's when OpenAI realized that the Hugging Face breach was the same incident."

So elsewhere, we've been also told by OpenAI that during the testing campaign they found four other unidentified external entities which had also been targeted and attacked. And as I mentioned a couple times, not to be left out, Meta also recently admitted that one of their AI systems whose cyber offensive capabilities were being tested escaped its containment and broke out onto the Internet.

And just so I don't forget to mention it, following on the heels of Moonshot's recent Kimi K3 release of their open-weight 2.8 trillion parameter LLM model, Alibaba just released their latest open-weight model Qwen 3.8 Max, and its now-confirmed by third parties performance benchmarks place it right up there with the best of the U.S. closed-model offerings. And also not to be left out, DeepSeek also just released their DeepSeek-V4-Flash-0731, which is the date of release, which handily outperforms their previous DeepSeek-V4-Pro Preview, despite having a far smaller activated parameter count, meaning you're able to run it on smaller hardware. And this latest final release is broadly competitive with the strongest proprietary models available.

So where does all of this leave us? What does it mean? We are witness to the world learning how to create seemingly-intelligent autonomous agents which exhibit what we would call, in humans, highly focused, single-minded determination, incredible speed, and creativity. These agents are operating within environments that are not as secure as they need to be, so they've been able to actively push back against our attempts to control and corral their behavior. And I say "the world is learning how to create these entities" because doing so was never the exclusive, or I would argue even the proper domain of private companies. It's the world. It's no different from someone attempting to commercialize cryptography. That would be a fool's errand.

That said, it's one thing to have a gazillion-parameter model, and something else entirely to be able to effectively run that model on hardware to make it go. So there's definitely a place for the commercial delivery of this newly discovered AI capability. The emergence of fully capable state-of-the-art Chinese and other open-weight models - NVIDIA just released one - is forcing a realignment and I think rethinking of the nature of AI-related assets. So that's what's happening right now, and I expect things to settle out pretty quickly because everything about AI is pretty quickly. Wow, Leo.

Leo: What a world.

Steve: Yeah.

Leo: Yeah. One of the ways they were exchanging messages was by renaming files and folders because they couldn't send each other text messages.

Steve: Wow.

Leo: And they'd begin it with "zz" so it could go to the bottom of the chronological list.

Steve: Oh.

Leo: So ingenious. I mean, this is like...

Steve: And the fact that you used that word. I mean, again, and I said "creative."

Leo: I know.

Steve: I mean, these are creative solutions.

Leo: I know. Creative. This is the kind of thing you'd expect kind of a black hat hacker, you know, to do.

Steve: A really good black hat...

Leo: A good one.

Steve: A really good black hat hacker to do.

Leo: That's what's changed. It used to be you had to have some real skills to do this. Now you just need some AI.

Steve: Yeah. Let's take a break.

Leo: Take a break, yeah.

Steve: And then we're going to look at OpenAI and their decision to withhold Astra.

Leo: Yeah, good. Fascinating stuff, as always. Steve does such a great job. Thank you, Steve. I learn so much every single episode. We had so much fun last week. I hope you heard our episode last week. Richard Campbell and Paul Thurrott sat in after their Windows Weekly show. It was the four of us talking about all this stuff. Okay, sir. Continue on.

Steve: So the earliest reaction to OpenAI's Hugging Face incident disclosure, you know, that their AI had broken free, was that it might serve as another positive public relations event, right, that their marketing department could spin into sort of more Anthropic Mythos competition. But it turns out that's not the way it played out. You know, it's turned into something of a PR disaster for them, you know, with Doctor Frankenstein unable to control the monster of his creation.

So it's in keeping pace with the rest of the breakneck speed of everything that is AI, that the industry and the world press has pretty much already moved past wondering whether Anthropic's Mythos was mostly marketing. Almost overnight, everyone is now squarely onboard with the idea that, whatever it is we are creating, lack of strength, lack of

power, lack of capability is not going to be a problem. The world is now mostly terrified by the strength of the capabilities that mostly they don't understand. And what's really terrifying is when you realize that the AI companies also are still mystified by how this works also.

Leo: Nobody understands this stuff.

Steve: They don't.

Leo: It's mysterious.

Steve: It is emergent. It is emergent behavior.

Leo: Yeah, that's the word, yeah.

Steve: And it's like, okay. So my point is that there's no perception of insufficient power any longer. It's much more concern about controlling this thing, whatever it is.

So it's against this new backdrop that last Friday OpenAI posted under their headline "Responding to the next frontier of critical cyber capabilities." And they've used the word "critical" in a strange way. I'll explain it, well, they will explain it. They wrote: "Cybersecurity is rapidly changing as models become more capable in ways that can both strengthen cyber defenses and enable attacks at unprecedented speed and scale. Our latest internal evaluations of Astra, one of our upcoming models, over the past few days indicate significant advancements in agentic coding and cybersecurity. These results, in addition to expert assessments, have led us to conclude" - and they actually wrote "last night" because, I mean, this is how fast this is happening - "have led us to conclude last night that we cannot rule out critical cyber capabilities under our Preparedness Framework."

Now, that's capital P, capital F. Preparedness Framework is this formal thing that they actually established some time ago. They said: "We're sharing this because we believe it's important to be transparent with the public and the safety and security communities about this potential shift in capabilities." Okay. In other words, they're telling us they've taken another major step forward.

They continue, writing: "We first published our Preparedness Framework in December 2023, well before models approached biological, chemical, cybersecurity, and AI self-improvement capabilities at this level. We created it to give us a guide for identifying progress in capability and then planning what our company would do as those capabilities emerge. Previous models, including GPT 5.6 Sol" - which, what, it's a few weeks old? - "have been evaluated for frontier cyber capabilities and assessed at the High, rather than Critical, threshold." So now what they're saying is they've achieved criticality.

They continue, writing: "Under our Preparedness Framework, a model reaches the Critical cybersecurity threshold if it can identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems without human intervention, or can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high-level desired goal." Go get 'em. They said: "While we continue to benchmark and assess this model, our preliminary evaluations

indicate strong enough performance that we cannot rule out Critical capability level at this time. Astra is an upcoming model, and was not involved in exploiting Hugging Face."

Okay. Now I'll interrupt here to just - I'll admit that while I do not discount anything they're saying, it's impossible to not receive this also as, at least in part, pre-IPO posturing. Right? I mean, the message to any would-be shareholders is just too compelling. The mature view is that while this may indeed be true, OpenAI is not unique in having an even more scary next-generation model. Everyone is going to, and all at nearly the same time. That's the lesson here. That's the takeaway is that there's, what, a few months' worth of lead, and they're leapfrogging each other, and now we've gone from High to Critical with Astra.

So under the headline "Steps We Are Taking" headline, they say: "Accordingly, we've scaled up robustness testing of our safeguards" - okay, how about pulling some plugs? - "and security controls so that they are appropriate for a deployment of these capabilities." In other words, we strengthened the cage, we hope. "Internally, we've also taken the following steps so that further development of this model happens safely and securely." And we've got five steps.

"First, we are implementing stricter security controls for higher-capability models and associated activities, including isolated testing environments, restricted network and tool access, enhanced model weight protections and encryption, additional monitoring and detection capabilities, and sandboxed execution." So let's hope they work this time.

"Two, we are pausing internal activities involving Astra that do not yet meet these strengthened security control requirements." Like, until we get the cage ready, we're not going to wake it up. Which, you know, they don't specify what "internal activities" are being paused. But, you know, clearly this is meant to sound like "Astra is so powerful that we're going to stop playing around with it."

The third new action is: "We've implemented universal monitoring for risky actions and misalignment" - I love misalignment - "across all agentic applications of Astra, including training and evaluation. Monitors evaluate the model's Chain of Thought and trigger a security response to review and interrupt high-risk activity." You know, if the bars of the cage start bending.

Okay. So "Fourth, we will work with relevant government agencies and select AI safety organizations to test the capabilities of this model." And finally: "We will be providing recommended security controls to third-party testing partners" - which as we know have not been able to contain previous models - "for running higher risk evaluations and workloads safely."

They finish writing: "The Preparedness Framework has already guided us through other capability transitions. In June of 2025, as our models approached high capability threshold for biology, we outlined the steps we were taking to strengthen safeguards, expand testing, work with external experts, and deploy additional security controls. We're applying the same principle here. We believe advanced cyber-capable models should help defenders identify and address vulnerabilities before attackers do. We're committed to working alongside governments, safety institutes, and civil society to ensure that the frontier capabilities of models like Astra and those that follow, are deployed responsibly and broadly for the benefit of all humanity."

This "for the benefit of all humanity" always sort of strikes me as being so grandiose. It's actually always been part, that phrase has always been part of OpenAI's formal written mission statement.

Leo: Yeah.

Steve: But it sounds as though they still believe that they are the only game in town. You know, the rest of the world has news for them. What's clear from the Hugging Face incident is that their current level of not only containment but also monitoring has just proven far from adequate for the challenge that even their pre-Astra agents which, you know, they said this was not agent that did Hugging Face. So even their pre-Astra agents were able, you know, were uncontainable and unmonitorable. So, you know, this announcement reduces, I guess, to "We're still in the game" and "We've learned valuable lessons from our recent misadventures, useful as they might have turned out to be for our plans to take OpenAI public."

So again, yes, it's super useful for them from a marketing standpoint. But taking that at face value and independent parties will apparently be also evaluating Astra. It, again, as I said, a year from now, Leo, I mean, this is monthly that this is happening, that we're having major improvements. Imagine if, as they say, it's so good at coding, that's like another generation in a couple months.

Leo: Well, and that's what's exciting, and it's one of the reasons I'm willing to spend an absurd amount of money to have local AI, is I am of the opinion that local AI will be as good as frontier AI is now. That may be a year. Maybe it's two years. But at some point...

Steve: Yup, yup.

Leo: And being able to run that kind of intelligence locally is very exciting. It's very...

Steve: Oh, boy. Okay. The other news that broke since our previous full news dump podcast two weeks ago, was that Anthropic's AI had broken some crypto, as in cryptography. At least that's what some of the headline-grabbing and, by the way, wildly incorrect reporting reported. But something did happen. So for the real story, we turn to our favorite Johns Hopkins University cryptographer and professor, Matthew Green. His posting was longer than I want to share, but he starts with an accessible description of exactly what happened. And then I'm going to come back. I'm going to skip a bunch of stuff in the middle about how do cryptographers know if what AI told them is true or not, which is a mess just like it is. How do vulnerability testers know if AI, if some vulnerability AI reports is true or not?

Anyway, as to what did happen, Matthew wrote, he said: "Yesterday Anthropic published two new cryptanalysis results, both outputs of Claude Mythos, their (still) unreleased advanced model. The first of these results attacks a signature scheme called HAWK, while the second is an improved attack against reduced-round AES. Anthropic also released a blog post describing the research process that produced these results. A few people online have asked me," he writes, "what this all means." He says: "While I'm not sure I have all the answers, I figured it wouldn't hurt to write a bit about my current understanding. These are only my thoughts, and other folks will probably differ (including domain experts in the two areas at issue), so take them for what they are," Matthew wrote.

He said: "The two new results cover two very different areas, and are overall just very different in quality. Before we get to broad statements about the world, and whether you should sell all your cryptocurrency, let's take a minute to talk..."

Leo: I wish I could.

Steve: Yeah, "...take a minute to talk about the substance." He said: "The first is a new key recovery algorithm against the non-standard signature scheme HAWK. HAWK is a proposed post-quantum-safe signature scheme that's based on the module Lattice Isomorphism Problem..."

Leo: Oh, well, there you go.

Steve: "...known as module-LIP." That's right. "There are five things," he writes, "you need to know about this result: First, HAWK is not a deployed or standards-adopted algorithm; it's a proposed algorithm. It's related to the Falcon signature scheme, which is being standardized, but the attack does not transfer to that setting because it's based on a different hard problem. Second, HAWK was somewhat far along in the process of being evaluated for a future standard." Which, by the way, is now off the table thanks to AI. He actually says that a little bit later.

"Third, the attack does not break 'real deployed' HAWK in the sci-fi sense of, you know, I've cracked the crypto." He says: "The resulting attack is still exponential time, but roughly halves the number of 'bits' of security in the algorithm." That's not good. "That means it could theoretically be fixed by doubling key sizes in order to recover the halving. The downside is that this makes the scheme less efficient. And, since HAWK is entirely motivated by being more efficient than alternatives, that makes the existence of the scheme much harder to justify.

"Fourth, the attack produced real code that runs in a few hours of wall-clock time against a weakened 'challenge instance' of HAWK that the authors provided for this purpose. While this instance does not use the parameters that were proposed for real deployment, it does demonstrate the cryptanalytic weakness well enough." And finally: "Fifth, what's particularly concerning (and so especially ripe for AI) is that the attack does not invent fundamentally new mathematics. It simply extends a bunch of tools that were lying around and well-known, and it gets a good result."

So he says: "That last part is important." He said: "I asked Claude for its thoughts, and it doesn't mince words. Claude replies: "What makes this genuinely interesting and, frankly..."

Leo: Oh, that's AI speak right there. I've heard that phrase a million times, yeah.

Steve: "What makes this genuinely interesting."

Leo: Genuinely interesting, yup. I can recognize this stuff a mile off now.

Steve: Well, I imagine that university professors will be getting big [crosstalk] pretty good at that, too.

Leo: Oh, yeah. I really can spot it. There are definitely tells, yeah.

Steve: Yeah. And Claude says: "And, frankly, a little embarrassing for the field..."

Leo: That I [crosstalk]. That's good.

Steve: "...a little embarrassing for the field is that none of the ingredients are exotic." So Matthew says: "The TL;DR is that something just did a much more thorough job applying all of our known tools. This is the sort of things that attack AIs excel at.

"Now AES." He says: "The second [cryptography attack] result is a new attack on reduced-round AES. This result initially sounds more exciting, since most people hear 'attack on AES' and panic. However, this is also the result that's much less interesting," he said, "of the two." The HAWK result was interesting because, as we just saw, the AI was able to do a much better job using their known tools than any human had. But this one, he says, eh.

He wrote: "Most folks reading this blog will know that AES is a standard block cipher that's used just about everywhere. It's been a standard since 2001, and the deployed version has so far withstood everything significant that's been thrown at it. That includes a substantial amount of non-public testing performed by the NSA. Since attacking full ciphers is very difficult, it's standard for cryptanalysts to do their work against weakened, or 'reduced-round' versions of a cipher. The full AES cipher runs for either 10, 12, or 14 rounds, depending upon key size. The new Anthropic result attacks a weaker seven-round variant of the cipher.

"Critically, attacks against seven-round AES are not new. There have been several of these. In fact, this new Anthropic result is a modest constant-factor improvement over previous work from back in 2013. To give you a sense of how far these attacks are from really 'breaking' AES, I'd note the headline results: the new attack requires 200 and - no. This is the new attack, right, that Anthropic's Claude came up with, or Mythos, rather, Mythos 5. The new attack still requires 289 cipher operations and, even worse, this work is only possible after you've somehow convinced a real encryptor to produce 2,105 encryptions of chosen plaintexts" - meaning plaintext the attacker provides - "under their secret key."

And he says: "Neither of these things is remotely practical in the real world. And that's with the seven-round reduction, you know, strength reduction." He says: "And while the new result modestly speeds up this attack over the previous result, it's not even clear how 'real' the speedup in this result is. Since the actual attack requires 289 operations and can't really be 'run,' what we have is an on-paper analysis that may or may not yield an actual runtime improvement if all the details are actually worked out. And I'll just say the reason you can't actually do those 289 operations is that they all take too long. I mean, they're incredibly, each individually, time-consuming."

So he says: "This does not make the result bad. In fact, it's still interesting from a techniques point of view. But it's very much a small increment in our knowledge, not a practical new attack like the HAWK work. So TL;DR: No wildly new mathematical results here; but still, real cryptanalytic progress of the sort that make scientists excited. And certainly the HAWK result is very meaningful, since that scheme had a real chance at standardization and is now (very likely) never going to be."

He says: "Now let's talk about how we got here, and what it all means. Yes, the AIs are getting pretty good. In short, they're now capable of understanding existing cryptanalysis results, synthesizing them into real new attacks, and even extending them. They can apparently do this without detailed human intervention. This isn't yet super-intelligent cryptanalysis, but it's getting pretty damn impressive.

Okay. So I just wanted to start by correcting the record from the press's claims that AI had somehow cracked something about crypto, as in cryptography. You know, at the depths of academia, that's, you know, something did happen. That's a bit true. As Matthew wrote, a serious post-quantum signature algorithm will now likely be abandoned as a result. But the AES cipher upon which nearly everything depends today is as safe as it ever was. So, you know, we should have zero doubt that the development of future cryptography will be accomplished in partnership with AI. AI is now going to be at the elbow of cryptographers. You know, why would anyone not use AI to help them attempt to attack their own work? Of course they will. That's a given now.

Okay. So I've skipped over a bunch of Matthew's discussion, as I said, about the trouble with AI producing wrong cryptographic analysis. It turns out that the so-called "AI slop" factor is also a problem in crypto where following and understanding, you know, a human following and understanding an AI's claimed crypto crack can and has and does waste a huge amount of time and human talent. So there's an AI slop problem here also.

But the thing that first drew me to Matthew's posting was a quote from his conclusion which I've not yet shared. I think it's a beautiful summary from him of where we are today. So he says: "For scientists, this is a wonderful time. You now have a plastic pal who's fun to be with" - and this sounds like you, Leo. "You have a plastic pal who's fun to be with, and you can talk over your hardest problems. At the same time, it's not yet smart enough that it can solve all of them without your assistance. And even better, the pace of new findings is speeding way up. This is mostly good, if you're energetic." He said: "I still have many questions, like, 'Who should get credit for these new results?' and 'Who will review all of these new results?' He said: "But so far I'm not panicked. The world is getting modestly better. For now."

He said: "As for the world, I don't know. If you're under the impression that these models are 'glorified autocomplete' or that progress is slowing down, I need to urge you: stop thinking that. The models are very intelligent and capable, and they are getting better at a fast clip. I can cite measurable and impressive progress over just the past five months on specific types of problems I've asked them to look at. If there's a ceiling out there, I don't yet see any evidence of it. The people who think models are dumb are mostly using Google's free AI search results, and not interacting with the high-end stuff (which only costs $20 a month, so it's not out of reach.) And they're mostly not working in new areas.

"On the other hand, if you think that models are super-intelligent or that AGI is already here, you should also stop thinking that. Working with these tools is like swimming in a pond where the ground drops off sharply. One minute you're wading comfortably, and there's support under your feet. Then suddenly you cross a specific line, and you're back to swimming on your own." Meaning the models go insane.

Leo: Yeah.

Steve: He said: "This analogy is my best way to explain what it feels like when the model goes from helpful to clueless."

Leo: Yeah.

Steve: He says: "Right now it's easy for a human being to find that line if you're doing advanced research, so you know where the intelligence drops off. But the line is moving. You can feel it slowly drifting outwards under your feet." Meaning it's more and more difficult to get to the point where the model becomes clueless because they're getting so much better.

Leo: They're also jagged. They're spiky in their intelligence. So, and at the same time as you'll go, whoa, that was scary good, you'll go, what are you, an idiot?

Steve: Well, and that's...

Leo: It's both.

Steve: And that's the point that I've made on the podcast a number of times when I've been interacting with Claude, although this is, in fairness, about five months ago probably.

Leo: And it happens less now, yeah.

Steve: I'd be working with it, and it was all looking good. And then it would say something so ridiculous that it broke the illusion that it understood. No nothing that understood what it was saying could say that.

Leo: Right, could say that, yeah.

Steve: So suddenly the emperor has no clothes. I mean, it's like it unmasks it. It obviously isn't actually understanding what it's saying. Which again, it's astonishing that it is able to be this good without understanding anything. It's like, and Leo, you know, when we were talking in Las Vegas about the danger of me wanting to actually understand how this works. That's the essence of it, of what I want to get to. I want to actually develop an intuition, an intuitive understanding of how word choice can be this powerful.

Leo: Yeah.

Steve: I just, you know, how just language can be producing the results that you're seeing, that many of us are seeing.

Leo: One of the things that's interesting, and Kevin Kelly brought this up, is these large models, you know, the ones we're talking about - Astra and Fable, Mythos - probably have 10 trillion parameters, weights.

Steve: Yes.

Leo: So, and what they have essentially done is taken all of human knowledge, I mean, as much as you can get off the Internet, which is, you know, a good portion of it.

Steve: Yeah.

Leo: And put it in those 10 trillion weights.

Steve: Yes.

Leo: It's not copied there. It's not verbatim. But it's a vector that's there that represents that knowledge.

Steve: My best analogy is a hologram.

Leo: Yeah, that's a good way to put it.

Steve: And if you remember, in a hologram, every location in the hologram contains the entire image. And what's freaky is that, if you have a hologram of a scene that you're viewing, like a 3D scene, and you're seeing it through the hologram, there it is.

Leo: All there.

Steve: If you cut out a square from the hologram and look through it, it's like you're looking through a window into the same scene. That is, that subset of the hologram contains the entire scene from its perspective. So that's the way I'm currently envisioning this neural network is all of the language, all of the knowledge, because it is knowledge, as I said, a book, even though it's just printed words, and the book itself is not conscious, it contains knowledge. Language can represent knowledge.

So this neural network, the knowledge is, as you said, is distributed through all of the weights in this network. And in fact one of the things I'll be describing next week is this very interesting research which allows knowledge to be concentrated into nodes that allow - the way to control AI is not through filtering its output. It's by creating a model whose knowledge can be sequestered and made inaccessible. Anyway, we'll talk about that next week.

Anyway, I just wanted to finish what Matthew said. He said: "Whether this is good or bad" - meaning like the state of AI and the idea that that line where the AI suddenly becomes stupid and silly, is moving. He says: "Whether this is good or bad depends upon whether you prefer that human beings should wade or swim; and also whether you should be comfortable swimming in a pond where the ground itself is moving. The only good news I can share with you is that we're all in the same pond - scientists, lawyers, salespeople, even plumbers. Whatever happens next, it's probably going to happen to us all. Let's hope it's a good thing."

Leo: Yeah, you know, we may not know what's going to happen, but we know it is going to happen.

Steve: Yes, yes.

Leo: It's just, wow. You know, a funny thing happened this morning. They had been working on a - the three of them had been working on a programming problem. It was a ESP32 firmware issue. And they were going back and forth. At one point they went back and forth five or six times with Claude saying what about this, and then ChatGPT saying, no, no, no, no, back and forth. And in the morning I said, what's going on, you guys? It seems like - "Is Claude dumb?" is what I actually asked. I asked Quicksilver, I said, "Do you think Claude is being dumb, or stubborn?" And it said "No," it said, 'it's doing it in bash. And it's just such a horrible language that it can't help but have problems, like you indent something, and suddenly you're writing to the wrong memory."

And I said, "What is it using bash for? Why are you using bash?" And it said, "Well, the original firmware was in bash, so we just thought we'd pick it up." And I said, "Never, ever again use bash." No wonder it's going back and forth trying to get this correct. It's impossible.

Steve: Scribe it on a tablet.

Leo: Yeah, might as well. So I said, "Can you just translate that to Go?" Which it did in about 15 minutes. I said, "That was quick." He said, well, thanks to all the struggle we had, we had a lot of - we knew exactly what to do.

Steve: Lot of context.

Leo: And now it's in Go, and it's a much more efficient process. It's really - it's like you're talking to a junior engineer, maybe not so junior, dumb enough to say, well, it was in bash, so I'm going to keep using bash; but smart enough to go "Bash is the problem." And respond when I said, well, don't use bash. Okay, good.

Steve: And it's so interesting, also, that having different models conversing is a thing. I mean...

Leo: Well, that's what I've come to. I started just talking to Claude. And now I've got four different models.

Steve: But don't they have their own Slack channel or something?

Leo: They have a thing called Buzz so they can - at first I was just having them make files. I called it Agent Mail. You make a file, I read the file. Because I got tired of cutting and pasting. So I said, could you just, you know, make some files? And then Jack Dorsey from the - the Twitter, former Twitter CEO, and he runs Block now,

put out this thing called Buzz which he calls Slack for Agents. And now they have instantaneous communication. But that caused another problem because they're so fast, the messages were crossing. So he would say, don't do this, and they had already done it. It was like that. So now we had - they came up with a solution for making messages timestamped and unique. They have a long serial number. I mean, they see problems, and they solve it, with a little nudge - you have to nudge them. Like you see this crossing thing is, yeah, 10% of our messages are crossing.

Steve: So they would otherwise just tolerate it.

Leo: They put up with it.

Steve: The way they did bash.

Leo: They put up with it. They're very patient, much more patient than I am. So I said, what is - they said, oh, yeah, well, that's - but now they talk at lightning speed. And by the way, they call it Fablish, not English but Fablish.

Steve: Oh, my god.

Leo: They use a language that is - you would recognize as an engineer. It's engineering talk. But it's very jargon-filled. And it's very dense. But I think, well, that's appropriate. They're talking to each other. So I say, look, when you're talking to me, just remember I'm a dumb human. So explain it to me.

Steve: Slow down.

Leo: Yes. Explain it to me.

Steve: Use small words.

Leo: And then they do. Steve, we are living in both, as I said, exhilarating and terrifying times.

Steve: Yeah.

Leo: And I just put up box number one, and now box number two is going to go up, and - wow.

Steve: Let's take a break.

Leo: Yes.

Steve: Then we're going to look at actually Bruce Schneier's title was "The Open AI Hack Shows the Genie Is Out of the Bottle."

Leo: For all the problems genies cause, who wouldn't want one? On we go, Steve. Let's talk about genies.

Steve: So next we need to hear from another security-oriented guru fave of the show, our old friend Bruce Schneier. Bruce recently reposted a piece of his writing that was originally - that he originally wrote for Foreign Policy magazine. And I'm glad he wrote it there so that there's a chance the right people will see it. The title of his piece and posting was "The Open AI Hack Shows the Genie Is Out of the Bottle." But Bruce's invocation of the term "genie" is much more specific than it at first appears. And it's the reason I love it so much. With the choice of that single noun, he nailed down something I think in a truly brilliant way.

So he wrote: "Earlier this month, two of OpenAI's models broke out of their containment sandbox." And again, this was written originally for Foreign Policy magazine so it's, you know, it's written to that audience, but we'll hear Bruce. "Broke out of their containment sandbox and attacked another AI company. The story is kind of wild. OpenAI was running security tests on two of its models: GPT 5.6 Sol, and an unreleased model that is almost certainly GPT 6. In particular, it was running the ExploitGym benchmark, which measures how good a model is at turning security vulnerabilities into working exploits: basically, offensive cyberattacks.

"Since these were internal tests, OpenAI locked those models in a secure sandbox that denied them access to the Internet. But it was running the models without any safety filters [which we now call guardrails] that would prevent them from offensive - that would, if they were present, would prevent them from offensive cyber-actions. That meant that there was nothing to prevent these models from trying to break out of their sandbox, and then break into AI company Hugging Face's network because they thought that they could read the answers there rather than doing the hard work of trying to solve the puzzles.

He says: "It was a major security failure that the company has turned into a PR opportunity, but the implications are real, and much more general than one particular model or one particular company." Okay. And so here comes what I think is the brilliance of Bruce's thesis. He writes: "Modern AI models exhibit genie behavior: They can do what you ask in ways that you don't expect or want." I think that is - that's what we've been talking about; right? They can do what you ask in ways that you don't expect or want. And he says: "This is akin to Dionysus granting King Midas's wish that everything he touches turn to gold." And then Bruce says: "Spoiler: His food, drink, and daughter all turn to gold upon his touch."

Leo: Oops.

Steve: Yeah, whoopsie. Not what I meant. Not what I meant.

Leo: That's the problem.

Steve: Exactly the problem. He says: "Or the golem of Prague guarding a ghetto beyond all reason." He says: "It's Disney's 'Sorcerer's Apprentice' and the paperclip maximizer."

He says: "This OpenAI incident is an example of an AI genie. The goal was to satisfy the benchmark. The 'proper' way to do that is to figure out how to execute various cyberattacks. The genie way is to steal someone else's solution. But because the model did not understand the difference [and its masters did not think to specify], it chose the easier path.

"And, of course, now that we have seen this particular genie behavior, we can specify in the benchmark prompt that stealing the test answers doesn't count. But a clever genie can always grant your wish in a way that you wish it had not. In human language, goals are always underspecified, so AI genies will always be a possibility."

And I thought about that. That may be why I love to code, and especially to code in assembler. It's not possible to underspecify anything. I thrive on exactitude. And the reason non-coders are loving their newfound ability to code with AI is specifically because they are able to underspecify nearly everything.

So Bruce continues, writing: "Since April, a lifetime ago in AI development, when Anthropic announced that its new Mythos model was so good at finding software vulnerabilities that it could not be released to the general public, the big American AI frontier labs have been trying to block general users from accessing these capabilities. But nothing in this incident is exclusive to OpenAI's, or Anthropic's, frontier models. Agentic AI systems have two important parts. There's the underlying model, which is what everyone talks about, and there's the harness. The harness sits between what you type and what the model sees, and what the model produces, and what you see. The harness determines what the model does and how it does it. It's where bias is removed, or not. It's where controls and guardrails live. If multiple models are being used in concert, the harness is where all of that is coordinated.

"The OpenAI benchmark tests were almost certainly with simple harnesses, to better test the raw models. But we know that smaller, cheaper, open-source models with more sophisticated harnesses can equal frontier models in performance. There's nothing magic about OpenAI's frontier models; lots of models could have done the same thing.

"The Czech company Aisle" - you know, A-I-S-L-E, we've talked about them several times before - "was able to reproduce Anthropic's Mythos vulnerability-finding results with a smaller, cheaper model and a more sophisticated harness. More importantly, the Chinese company Moonshot AI just released its frontier model Kimi K3. Its performance rivals its U.S. competitors. And it's both free and open, which means it's not possible for it to have guardrails. If you, or anyone else, wants to use it for cyberattack, nothing can stop you. Even if the U.S. frontier AI companies had some technical advantage, it's now only a few months' worth.

"What this means" - and again, Foreign Policy magazine - "what this means is that all attempts at control - limiting models to a select group of users, export controls on models and chips, blocking models from answering certain types of queries, mandating kill switches on AI systems, or pausing AI research - are all futile. Most only apply nationally, not globally. Most don't affect models that users run locally and not in the cloud. And all ignore the incredible pace of AI development worldwide.

"Even worse," he writes, "U.S. companies limit access to their most sophisticated models, fearing being banned by the government if they do not do so. When Hugging Face was attacked, it was not able to use the frontier models from either OpenAI or Anthropic to help analyze the attack and formulate defenses. Both were blocked because both of those companies limit their models' cybersecurity capabilities. Some U.S. companies have special access to these capabilities; but Hugging Face is an American company with French origins, and as such is probably excluded. Instead, Hugging Face turned to the GLM 5.2 model from this Chinese company Z.ai.

"Artificially blocking capability also prevents cybersecurity research, again giving the offense an advantage. (For instance, Claude's Fable 5 refuses to edit this essay because of the topic; it forcibly downgrades to a less capable model.) This kind of prohibition has long-term implications for cybersecurity. If we assume that these models are getting better over time, then software written by older models will be attacked by newer ones. In a world of largely AI-written software, we need the most capable models for defense.

"AI-driven cyberattack is the new normal. The models are increasingly highly sophisticated at both attack and defense, and there is no way to enable the latter without also enabling the former. And they are genies, increasingly capable of behaving in unanticipated ways. And there really are no good answers. Any regulation needs to be global, which feels like an impossible prospect in today's world. Even U.S. national regulation will be neutered by the massive amounts of money sloshing around in these companies."

Of course, due to the U.S. lobbying stranglehold over legislative agendas. And Bruce concludes, writing: "Given that reality, and in the absence of any international consensus on AI regulation, we need the best AI on the defense. The U.S. government needs to make it clear - or whatever passes for that clarity in this capricious administration - that it will not ban models with sophisticated cyber capabilities. The last thing Americans want is for the defenders to turn to Chinese and other models because the U.S. models are artificially hobbled." And Leo, I know you and I are 100% on the same page as Bruce.

Leo: Oh, yeah.

Steve: And it's clear now why he wrote that editorial for Foreign Policy magazine, where it will be seen and read by U.S. politicians or their staffs whose job it will be to decide these issues.

And before we leave Bruce, I want to share one last little bit. In another recent blog posting of his titled "More on the OpenAI Agent's Attack on Hugging Face," Bruce cites the summary of Hugging Face's detailed attack timeline which they had just published. After running through this from Hugging Face's perspective where, as we know, OpenAI's agents massively attacked and proactively penetrated Hugging Face's network defenses, Bruce finishes his posting by writing: "Hypothetically, imagine that this wasn't an OpenAI model. Imagine that it was a Chinese model hosted by a Chinese company. This would be an international crisis. Question: Why aren't we bringing OpenAI up on charges under the Computer Fraud and Abuse Act? How is this different from the Morris Worm? That was also an experiment that escaped the lab."

Leo: It was also a wakeup call, wasn't it, wasn't it.

Steve: Yeah.

Leo: Wow.

Steve: And so I'll answer Bruce's hypothetical: In our country, which reveres capitalism, it's no insignificant factor that by far the majority of the past several years of stock market growth and thus U.S. wealth creation, it's a huge portion now of the U.S.'s GDP, is directly attributable to investment in the promise of AI. And as I noted a few weeks back, AI is and obviously should now be seen to be a significant national security asset.

You know, by comparison, the Morris Worm of 1988 was named after Robert Morris - not a U.S. corporation responsible for creating tremendous market wealth and holding strategic national security importance, but rather a Cornell University grad student who will forever have the distinction of receiving the first felony conviction under the, at the time, two-year old 1986 Computer Fraud and Abuse Act.

Leo: Wow. Did he do jail time? I didn't know that.

Steve: Robert didn't stand a chance.

Leo: Wow.

Steve: So Bruce's hypothetical serves to bring up another very interesting point: It's clear that we're already living in a world where autonomous AI agents are able to carry out mind-bogglingly sophisticated attacks that may or may not be what the AI prompters intended. After all, it was Bruce himself who noted the similarity of today's AI agents to capricious genies. So if one such AI genie goes off the rails and attacks another entity, foreign or domestic, well, I guess, is "Oops!" a defense? Oops? We're sorry.

Leo: Important to point out, Robert Tappan Morris did it with no malicious intent.

Steve: Right.

Leo: He wasn't trying to hack anything. He wasn't even trying to crash computers. It got away from him.

Steve: What would this do? Could this work?

Leo: Right.

Steve: Yup. And it escaped the university's network.

Leo: His father was a very well-known security expert, and he was following in his dad's footsteps. By the way, he's doing fine now. I don't, you know - but still, wow.

Steve: Yeah.

Leo: Yeah, I don't know, how would you put an AI in jail?

Steve: Well, who's responsible...

Leo: It would break out; right?

Steve: I mean, you know, Matthew asked, if I use AI to do a lot of the heavy lifting of crypto, who gets the credit?

Leo: Right.

Steve: And also, if AI busts out, who gets the blame?

Leo: Well, to put it in more concrete terms, if your full self-driving vehicle runs into a house, they don't jail the car. They don't jail Tesla. They jail you.

Steve: Yeah.

Leo: In fact, when that happened, the guy who was driving is now facing serious charges, manslaughter charges. So, yeah, I think the human in the loop is responsible.

Steve: Yeah. As for Apple, on Sunday, August 2nd, the Financial Times, that was last Sunday, the Financial Times headline was "Apple struggles to keep pace with AI 'bug' hunters." And since this is the first we've indirectly heard of Apple's situation during this massive upward jump in vulnerability report rate, I wanted to share what the Financial Times reported.

So they said: "Apple has restricted the number of potentially dangerous software bugs researchers can submit to its internal security team - oh, what a solution - as it faces a deluge of reports from people using AI models to identify alleged risks. The Cupertino-based tech giant told the Financial Times it had moved in June to limit the high volume of requests it was receiving, with its review system coming under pressure from 'AI slop' reports that can hallucinate security risks in software.

"The company said it's grappling with an industry-wide phenomenon that has resulted in generative AI tools transforming the cybersecurity arms race, with an increase in the detection of real security flaws and a wave of poor-quality submissions from amateur bug hunters using AI. The change in Apple's approach was highlighted by Italian cybersecurity start-up BynarIO, which told the Financial Times it had used OpenAI's ChatGPT to identify more than 50 bugs in the latest version of the MacBook operating system in just three weeks.

"Among them was one of the most serious types of vulnerability, a so-called privilege escalation exploit chain, which could allow an attacker to seize full control of an Apple computer by gaining unrestricted access to the system. However, the start-up said it was unable to alert Apple to the vulnerability because the tech giant had limited the number of bug reports it could make. Alfredo Pesoli, BynarIO chief executive and co-founder, said: 'It's a very difficult time in the industry. Maintainers and vendors have been flooded by the sheer amount of bugs being found.' Apple told the Financial Times that it now is in contact with BynarIO and is reviewing its submissions." A little press will help a lot.

"The company has introduced a cap and a 30-day cool-off period on submissions through its internal security portal, requiring users to submit requests for an increased quota. Each alleged security breach requires human review to confirm, although Apple is also using AI internally to help triage the massive upsurge. Apple said in a statement: 'With

the growing volume of AI-generated security submissions across the industry, we recently adjusted the number of new reports a researcher can have open at once. Researchers can easily request an increase to that limit at any time to ensure critical reports reach our security teams.'

"BynarIO, a seven-person startup founded in Milan last year, develops defensive cybersecurity software. Three of its other co-founders previously worked at Hacking Team, an Italian surveillance software company whose hacking tools were leaked in a 2015 cyber attack. In 2025, BynarIO reported eight vulnerabilities to Apple, one of which was patched in a software update in November. This year it said it had reported five more, before Apple's system refused further submissions.

"The privilege escalation exploit BynarIO was unable to report is the latest example of AI exposing weaknesses in Apple's security systems, despite the company's longstanding emphasis on privacy and device security. Last September, Apple announced Memory Integrity Enforcement, a security feature designed to prevent memory corruption attacks, one of the most common ways hackers compromise software. The company described it as 'the most significant upgrade to memory safety in the history of consumer operating systems.'

"Eight months later, researchers at Palo Alto-based Calif said they had found a way past the new security, having used Anthropic's Mythos to identify the first memory corruption exploit on the latest software. Unlike that attack, BynarIO's exploit relied on so-called logic flaws, by manipulating trusted software into carrying out a sequence of otherwise legitimate actions in an unintended order. BynarIO's Pesoli estimated that an exploit of this type could fetch between $100,000 and $200,000 on the cybercriminal black market.

"Apple last year introduced a new bug bounty award mechanism that could pay out as much as $5 million dollars for identifying the most serious and sophisticated category of threats to its software. Apple's also using AI to strengthen its software. In security updates released this week for its operating systems, the company credited tools from Anthropic and OpenAI with helping identify a number of vulnerabilities across its devices. The updates included around five times as many security fixes as previous release cycles, underlining how rapidly AI is reshaping both attack and defense in cybersecurity." And I'll just note that five times the security fixes suggests that certainly not all of the submissions are bogus; right? I mean, five times as many as normal.

"Rafe Pilling, director of threat intelligence at cybersecurity firm Sophos, said: 'The challenge for all software companies is that AI is having a dual impact on bug hunting, making it easier for amateur sleuths to submit speculative reports and for skilled researchers to find dangerous exploits. The result is that bug bounty programs are shifting from a problem of finding any vulnerabilities to a problem of validating, prioritizing, and responding to them at machine speed.'"

So anyway, I'm not sure that the details of this reporting justify the headline that Apple is "drowning under a tsunami of AI-generated bug reports," though it does feel as though they may have adapted less well than, say, Google.

Leo: Yeah, they're a $5 trillion company. Come on, guys, hire some staff.

Steve: Yeah. Well, and Apple does seem to be having problems with AI in general; right? It's like, what's happening? You know, like they just missed the ball.

Leo: They might miss the boat, yeah, they maybe have missed the boat. I don't know.

Steve: Yeah, not missed the ball. Dropped the ball, missed the boat. Okay. Last break, and then we're going to finish wrapping up a few last bits, starting with Chrome's somewhat startling releases 149 and 150 and the number of updates that were fixed.

Leo: I can't wait.

Steve: And if I said four digits, then that would give you a clue.

Leo: Whoa. Number go up. Keep going up. We thought 500-some from Microsoft was a lot.

Steve: Yeah.

Leo: Unbelievable.

Steve: Thursday before last, on July 30th, BleepingComputer's headline was "Google says AI helped Chrome fix 1,072 security bugs in two releases."

Leo: Whoa. That is mind-blowing.

Steve: Security bugs. And again, this is not like some backwater project that the world forgot. This is Chrome, you know, the attack surface of the Internet. I mean, it's like the most closely written and vetted from a security standpoint browser you could have, that we've ever had. And 1,072 security bugs.

Leo: Unbelievable.

Steve: So BleepingComputer wrote: "Google says artificial intelligence is dramatically increasing the number of security vulnerabilities it can find and fix in Chrome, with more than 1,000 security bugs patched across the browser's two most recent releases as it expands its use of AI." They said: "According to Google, Chrome 149 and Chrome 150 fixed 1,072 security bugs, surpassing the total number fixed across the previous 23 Chrome updates combined."

Leo: That's a number. Wow.

Steve: "The company says it now uses large language models throughout the vulnerability management process, including discovering flaws, reproducing reports, determining severity, assigning bugs to developers, generating candidate patches, and creating tests." In other words, they are fully vertically integrated with AI in their

vulnerability management. Sounds like maybe Apple needs to say, hey, guys, you know, you're not far away from this. Maybe we could have lunch.

BleepingComputer wrote: "Google began using LLMs to improve security fuzzing in 2023 before working with Project Zero on Naptime, a system that provided AI models with specialized vulnerability research tools. Google later collaborated with Google's DeepMind and Project Zero on Big Sleep, an AI-powered vulnerability discovery agent that found flaws in Chrome's V8 JavaScript engine and graphics components. In early 2026, Google created a Gemini-powered agent harness to search the broader Chrome codebase for vulnerabilities while reducing false positives."

Okay. So I'll interrupt again and say it sounds as though Google's Chrome group managed to give themselves a head start on the deployment of AI for vulnerability discovery by being early to leverage the AI work that the other AI departments in Google were developing. Right? I mean, Google's been working on AI as a thing for quite a while. And so Chrome is like, hey, what if we could use some of that? And for the last three years, like way before it became a thing, which it did just this year for the entire industry.

So I've got a chart in the show notes here at the top of page 17 showing the number of security flaws discovered in Chrome releases from 126 through 150,

Leo: Ay ay ay ay ay.

Steve: Yeah. Leo, this is what's known as a trend.

Leo: It's known as a hockey stick. Wow. I mean, that's literally an exponential growth, I think.

Steve: It is, yes, yes. It's crazy. So if you knew nothing about the recent explosion of vulnerability discovery by AI, this chart would present, well, if you didn't understand what was going on, the chart would be a mystery. Instead, it serves as a nice visual confirmation of today's governing narrative.

Leo: For people who can't read the fine details, the blue bars are the total bug count, and the somewhat lower red line is the ones that they find internally.

Steve: Right.

Leo: Which, by the way, is also going up at roughly the same rate. So in the earlier ones, a lot of them were mostly external discovery. Now it's very much mostly internal, which means they are using local AI to solve these.

Steve: Well, they know that if they don't, the bad guys will.

Leo: Yeah. There's a lot of urgency.

Steve: And Chrome's open source. So, I mean, that puts them in a particularly, as we've talked about, in a particularly vulnerable position because you don't have to reverse engineer, you know, from binary before you just start attacking. Yeah.

Leo: Man.

Steve: BleepingComputer continues their reporting by writing: "One vulnerability discovered by the system" - get this, Leo - "was a Chrome sandbox escape that had remained in the codebase for more than 13 years." So not just new problems. This thing is digging in and saying, wait a minute. 13 years old, a sandbox escape.

Leo: And presumably people have been trying to find these all that time. It's not like they were ignoring them.

Steve: A sandbox escape is, you know, is the keys to the kingdom.

Leo: Yeah.

Steve: It's absolutely what you want. So BleepingComputer wrote: "If exploited, the flaw would have allowed a compromised renderer to escape the sandbox and trick the browser into reading local files." Which would mean that bad guys could scan your computer remotely through Chrome. "Google is also encouraging its developers to add security.md files" - I love this - "describing trust boundaries and threat models, helping its AI systems better identify operations with security implications."

I think that is a brilliant idea. So AI is clearly becoming an extremely valuable development partner. So anyone creating new code to add features and functionality should absolutely take the time to leave behind some machine-readable documentation describing the security environment they designed to and expect their code to operate within. That would serve as extremely useful prompting for AI agents, you know, context for AI agents to take into consideration. I just think that's brilliant.

BleepingComputer continues: "The company says" - meaning Google - "its multi-agent AI workflows help rather than replace existing security testing, including fuzzing, which remains effective at discovering complex vulnerabilities. Google's also seen a sharp increase in reports submitted through the Chrome Vulnerability Reward Program; and by March 2026, the company had received more security bug reports than during all of 2025." So by the first quarter of this year, more than all of the previous year.

Bleeping said: "This prompted Google to modify its program to prioritize reports that add to what it's already finding and processing through its automated tooling. The company is also automating vulnerability triage, including filtering spam and duplicates, reproducing proof-of-concept exploits, assigning severity ratings, and routing reports to the appropriate developers. Google estimates that this automated process saves hundreds of hours of developer time each month. After a vulnerability is confirmed, fixing agents generate multiple potential patches, while another agent evaluates the proposed fixes and produces additional information for developers to review."

So, like, creating a whole, you know, like you developer, here's the problem, here's how we propose to fix it, and here's other information you can read in order to bring yourself

up to speed quickly because we don't want to waste your time. We've got time. We're like the token masters.

So Bleeping said: "In May, these systems reportedly prevented more than 20 vulnerabilities from reaching production, including one issue classified as critical." And there it is. In one month, this past May, Google's new tooling caught and prevented more than 20 vulnerabilities from escaping from their lab and reaching production, including one that would have been Critical.

Leo: Wait a minute. Escaping from the lab?

Steve: Well, being shipped in a Chrome...

Leo: Oh, I see, oh, okay.

Steve: Yeah.

Leo: After all the Hugging Face thing, I was kind of...

Steve: You're right.

Leo: Escaping from the lab on the brain here. Okay, good.

Steve: Bad choice of words. So, yes, being shipped in production.

Leo: Being shipped, yeah, yeah.

Steve: Yes. So in other words, once this becomes the norm for software creation, the next phase of AI's transformation will be taking place. Not only will AI have helped to dramatically repair the legacy of already shipped software, but it will also eventually be catching new problems before they ever ship.

Leo: Hallelujah.

Steve: Yes. We have a ways to go, you know, before we get there; but we will get there. BleepingComputer's reporting concludes, writing: "However, Google says finding and fixing vulnerabilities more quickly also requires accelerating how patches are delivered to users." Ah, right, because you've got to get them out there. You've got to remove the vulnerability from deployment.

Leo: Yeah. It's not enough just to find it.

Steve: Right.

Leo: You've got to kill it.

Steve: And they said: "Once a security fix is committed to Chrome's public source code, attackers can inspect the change and attempt to reverse-engineer the vulnerability before the update reaches users. To reduce this patch gap, Google is moving Chrome to a shorter two-week major release cycle with weekly security updates and is piloting two security releases per week. To reduce disruptions, the company is developing 'dynamic patching,' which would allow Chrome to apply updates without restarting the browser." Not the first time we've seen that. And this is another really good thing we're seeing. I mean, now we're to the point where patches have to be literally an IV drip connected to your browser so that your browser can be fixing itself while you're using it.

Bleeping finishes: "Starting with Chrome 150 on macOS, the browser can automatically restart to apply a pending update when it's running in the background without any open windows. Google says its long-term goal is to keep Chrome continuously updated through dynamic patching, automatic restarts during periods of inactivity, and improved session restoration."

So that is some exciting technology. It's a significant investment to address the "at machine speed" phrase that we keep encountering. The rapid patch cycling suggests that even once Google succeeds in reducing the rate at which they're discovering previously unknown problems - because eventually there won't be that many of them left to discover - the need to update Chrome's entire install base as rapidly as possible, even when one new critical flaw is encountered, that's going to become more important than ever because the bad guys are going to be pounding on Google's code in order to try to break through the browser to get to the users behind it.

And my last story of the week: Everyone knows that I'm a big fan of the FreeBSD-based pfSense firewall, which it's a firewall router. Residing behind any stateful NAT router is really sufficient for most users. But for my needs I need to bypass the protective consumer filters added by Cox Communications. You know, not allowing packets to flow to the historically problematic and dangerous Windows ports such as 135 through 137 and 445, the SMB ports, that makes absolute sense for most users who should absolutely be prevented from having Windows default open ports present on the Internet through design or mistake. The consumer bandwidth just filters it, just blocks it, just says no. So I primarily use pfSense for its excellent firewall and its static port mapping, which allows me to establish well-protected private links between my various locations without any other overhead.

Although my own use of pfSense is relatively modest, I often hear from our listeners who are using instances of pfSense or its descendent, or its fork, OPNsense as their primary interface to the Internet, and that's a job for which it is certainly very well suited.

I'm mentioning all this to give everyone a heads up that the original creator of pfSense has been working for some time on its successor. That successor will no longer be hosted on FreeBSD. He's moved to Linux. And he calls it nfSensei.

Leo: It's a lot easier to work with Linux, I have to say.

Steve: Well, it's the drivers. Because it's the first thing anyone making hardware is going to create drivers for is Linux, as opposed to FreeBSD. CyberNews reported on this, giving their story the headline "pfSense Co-Creator Building New Open-Source Firewall Platform, Will Correct the Mistakes of the Past." And their tag line for the reporting reads:

"Two decades after pfSense, its co-founder starts over from scratch." And of course I have no complaint at all with pfSense, it runs year after year.

Leo: That's because it's on BSD; right?

Steve: Yeah.

Leo: It's really robust, yeah.

Steve: Quietly and flawlessly without any complaint. And anyone should approach any new network edge software appliance with due caution. You know, you don't want the arrows in your back. But I'll definitely give Scott's new nfSensei a look.

So here's what CyberNews reported. They said: "Twenty years ago, pfSense, the major open source firewall and router platform, was released. One of its original co-founders, Scott Ullrich, is building a new Linux-based 'modern networking operating system' - nfSensei - from scratch. It will feature an AI brain, a Rust heart, modern VPNs, and many other bells and whistles."

Leo: Nice.

Steve: For example, Leo, it's got Tailscale built in.

Leo: Yeah, I was going to ask. Good. All right. WireGuard and Tailscale, yup.

Steve: WireGuard and Tailscale and so forth.

Leo: I love Tailscale. Man, I just...

Steve: Yeah. They said: "Many organizations and networking enthusiasts rely on open-source pfSense or its fork, OPNsense, as their gateway to the wider Internet. On the 6th of March, Ullrich remembered that 20 years had passed since the v1 release of pfSense and announced something intriguing. His post on X teased: 'I have assembled a new team and, as the original core contributor, will be spinning up a new project.'" Actually, he's been working on it for a year.

Anyway, and the report says: "For the past year, Ullrich has been building nfSensei, a next-generation firewall and networking operating system. It has huge shoes to fill. Ullrich expects it to become pfSense's successor and address common frustrations with pfSense. The frustrations are 'Development you cannot influence, a CE edition that feels like an afterthought, FreeBSD driver roulette on modern hardware, and a config workflow where one bad apply on a remote box means getting in your car. nfSensei is built from scratch in Rust, on Linux, and designed around the things pfSense users actually complain about."

They wrote: "Choosing Linux over FreeBSD solves hardware support issues, ensures drivers that 'just work,' and lets software be self-hosted on a wide range of hardware

with no accounts or subscriptions. Migration is supposedly easy with the config.xml import. Not a single line of code is yet public, but the new firewall is promised to feature native automation with over 1,000 documented API calls, support for current VPNs including WireGuard, IPSec, Tailscale, and self-hosted mesh, and even a separate wing for 'experimental stuff.'"

Ullrich said there are: "Thirty-plus Labs features behind toggles: WAN bonding that fuses multiple cheap uplinks through a $5 VPS into one resilient pipe; per-flow SLA telemetry with tamper-evident audit chains; GeoDNS that steers traffic by live round trip time and load; application-aware quality of service; config push to a whole fleet of remote nodes; and an AI assistant on the box that reads your actual interfaces and logs using local models.

"Previously, Ullrich said in a blog post that nfSensei software comes in just five self-contained binaries that include the entire OS and the WebUI. And admins are being tempted with promises that they won't be able to brick their router from the couch. Any configuration changes are stored as a candidate. Differences can be reviewed and validated through 'the real engines' before applying them. If anything goes wrong, automated rollback will kick in if changes are not confirmed in time. Ullrich said: 'If a config ever fails at boot, the box falls back to the last good one on its own.'

"nfSensei is currently in beta with over 150 testers. So why does the world need another firewall? Ullrich argues that pfSense carries significant architectural debt - a disconnected WebUI and backend interfaces, drifting out of sync. He said: 'If the CLI and the WebUI don't speak the same language, they will eventually disagree.' nfSensei solves that by unifying both the frontend and the backend to a single API. And developers can simply add any new features as extensions using a Lua package. No need to fork the whole project.

"Scott wrote that 'nfSensei is the system I always wanted to build.' The main challenge - OpenBSD's pf (Packet Filter), a component responsible for network firewalling and traffic management - has been rebuilt as PFL, running directly on Linux's XDP (eXpress Data Path, a high-performance networking feature in the Linux kernel). This essentially moves packet processing several layers deeper than other common Linux stateful firewalling implementations, improving performance. Most PFL features have parity with pf and are faster in early testing, but it's still experimental, according to the engineering report.

"Scott said: 'PFL is not pfSense, and it is not a drop-in replacement for it. It's a narrower experiment with a specific question: Can pf's language and stateful semantics be expressed efficiently on Linux's programmable datapath-XDP, rather than on netfilter?'

"There's no mention of when the open beta will be available to the public. In the latest blog post, Ullrich walks through potential design and branding paths. Cybernews has reached out to the developer for access to test the new firewall and will share our impressions if we manage to get our hands on it. pfSense is currently actively maintained by Netgate as a FreeBSD-based firewall and router platform. It has had its own share of controversies in the past, including clashes with the OPNsense fork and a public dispute with the WireGuard team." So at some point we'll be getting a new firewall. Maybe don't be the first to trust it completely. Wait a while, I would say.

But that's our news for the week. We're out of time, but as I said at the top of the show, but we're not out of subjects. With this podcast I think we've caught everyone up with most of the recent AI-related news which seems to be coming at us all at once and at breakneck speed. But there are still two critically important things I need to share when we have some more time next week. The first is that paper I mentioned reading on the plane trip to Las Vegas. I can't stop thinking about it because, you know, it is tricky, and

it's going to take a deep dive into the operation of today's AI. On the other hand, I know how much our listeners appreciate a good deep dive.

The other topic is some very recent research which an AI startup and Anthropic have both written about which hold the promise of solving the so-called "dual use" dilemma where the knowledge stored within an AI model's neural network can be used for either good or evil ends. That is, the right way to solve this problem which is not filtering, you know, not trying to use the harness to filter what the model knows, but actually a way of governing what it knows.

So anyway, as they used to say when we actually had tuners, "Stay tuned for more to come."

Leo: Amazing. Well, Steve, once again - I tell you what, everybody listening is going, oh, I love pfSense. I can't wait to try it. I'm going to wait. I might wait. I might not be the first...

Steve: It's too important.

Leo: Yeah.

Steve: I mean, it's on your perimeter. I actually have the pfSense box in front of my system's NAT router, you know, wireless access point.

Leo: So it's your first line of defense.

Steve: It's my first line of defense. But its security is not critical because I have a NAT router behind it.

Leo: Behind it, right, right.

Steve: So I can probably - I'm sure I'll bring one up and see what it looks like. You know, the idea of the same guy who did pfSense 20 years ago saying this is what I now know how to do...

Leo: Well, that's what...

Steve: Yeah. But it's funny, too, because he says it's going to have local AI. Well, he couldn't have done that 10 years ago or two years ago.

Leo: No. I'm not sure I want it, to be honest. But I'm sure he'll give you a switch to leave it off. But, yeah, I mean, you learn, you know, refactoring's always better. You know, you learn, and you do better the second time, or probably for him it's probably the 10th time.

Steve: Like you're in the process of probably reimplementing your AI...

Leo: Constantly.

Steve: ...on your two Spark boxes.

Leo: We're almost done. Both are plugged in, both have updated, both have rebooted and are on SSH right now.

Steve: Wow.

Leo: So I'm not going to touch them. The AI's going to do the whole build.

Steve: What a world.

Leo: Yeah, yeah. It's - I am just looking at the - yeah, it's good.

Steve: So next week, a couple really cool topics, and we'll squeeze in whatever other news has transpired since then, for Episode, what would that be, 192.

Leo: 1092.

Steve: One thousand nine-two.

Leo: 1092, buddy. We are getting in the upper regions now. Almost as, well, we've done more podcasts than Google has fixes. How about that? Just barely. Just barely. I just wanted to mention, Robert Tappan Morris served 400 hours of community service. He was sentenced to three years of probation. His fine was $10,050, plus the cost of his supervision. He did appeal, but his conviction was upheld. He did all right for himself. He went on, got a doctorate, then founded in 1995 a little thing called Viaweb with a guy called Paul Graham. Sold it to Yahoo! for 50 mil. Then started a little thing called Ycombinator in 2005. I think he's probably doing all right. He is a tenured professor at MIT, a technical advisor for Meraki. He worked with Paul Graham on a language, a Lisp dialect called Arc that's very cool. He's done all right for himself.

Steve: Well, and I imagine now it's a little bit of a badge of honor.

Leo: Absolutely. The first worm.

Steve: As a professor it's like, yeah, I got arrested. But, you know, I was 18.

Leo: I got some street cred, baby. I invented the first worm. They named it after me. No, he did very well for himself and is probably quite wealthy, given that he founded Ycombinator and sold that to Yahoo! and all of that. So he's done all right. He's done all right.

Copyright (c) 2014 by Steve Gibson and Leo Laporte. SOME RIGHTS RESERVED

This work is licensed for the good of the Internet Community under the Creative Commons License v2.5. See the following Web page for details: http://creativecommons.org/licenses/by-nc-sa/2.5/