Models Go Rogue and ExploitGym

Description: OpenAI's unconstrained internal testing AI got loose, attacked Hugging Face. We hear from OpenAI, Hugging Face, and Andrew Ng. GRC went off the air Friday. Was GRC hacked? What happened? The Linux kernel project repairs 442 CVEs in a single batch. LG's PC monitors cause PC adware installation. France bans all social media access below age 15. WordPress's recent Critical vulnerability claims victims. Amazing details about "Rocky" from Andy Weir. The new AI exploit ranking benchmark that caused the breakout.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1089.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1089-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. Big show. Big show. We're going to talk about that wild story of the OpenAI model that escaped containment and hacked Hugging Face. France banned social media access for kids under 15. WordPress has a critical vulnerability. And was GRC hacked? That and more coming up. Security Now! is next.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1089, recorded Tuesday, July 28th, 2026: "Models Go Rogue & ExploitGym."

It's time for Security Now!, yes, the show we wait all week for. Tuesday's here. And when Tuesday's here, so is Steve Gibson, the main man at Security Now!. Steve.

Steve Gibson: I do know that I, at least, wait all week for this, because, well, [crosstalk]...

Leo: You work all week for this.

Steve: ...until the next time it happens, yeah. I imagine other people...

Leo: Do you, at the end of Security Now!, do you breathe a sigh of relief and, well, that's over for another few days?

Steve: Yes, because it's the longest interval before the next one. So it's like, okay, that's behind me. So now I get to do work until I have to get ready for the next one. Although this last episode of July, for July 28th, is a little different because next week I will not be

mailing show notes for 1090 because you and I are going to be doing a different Security Now! from the ThreatLocker booth during the Black Hat event on Wednesday. My plan is, I ran across something regarding AI that stuck with me. So I'm planning to still do emailing to our subscribers about something I think is really interesting about this notion of AI alignment that is beginning to surface which suggests a way of getting AI not to misbehave by removing the knowledge that we don't want it to have, which is really interesting.

Leo: Ah, interesting, yeah.

Steve: Because, you know, how in 2.8 trillion parameters, like the knowledge, it's like holographically stored; right? It's like all the knowledge is everywhere. And it turns out there's a way of getting the knowledge to group into, like, a region, and then you excise it, kind of like a little tumor.

Leo: Oh, interesting.

Steve: Anyway, I'm going to - I'll be doing an emailing on the weekend. But then - oh, and I always wanted to tell our listeners, I think I mention at the end of the show, but I'll say it right now in case people don't listen all the way through, you and I are going to basically be having a conversation using talking points from the Security Now! mailbag feedback.

Leo: Nice.

Steve: As we're sitting in the booth. So I wanted to...

Leo: We used to do those feedback episodes. This will be a feedback episode.

Steve: Essentially, yeah. But so you and I will just take - actually, because it's Black Hat, I'm going to print them on paper rather than, you know, have any electronic device which is on. I mean, I can turn off all the radios, I realize. But anyway, why not have paper? And so you and I will just be using...

Leo: This'll be interesting.

Steve: ...thoughts from our listeners. So I wanted to solicit any talking points during the next week from our listeners who would like to hear some point addressed. That's sort of generally what I have anyway. And then it's like, it's not like I don't have plenty to work from. But I thought, okay, that'll be fun, just to say that's what's going to happen. So...

Leo: We should make it, if you're going to Black Hat, come by the ThreatLocker booth. But it isn't going to be audience seating kind of a thing like we did at ThreatLocker. It's just a booth. And I don't - I honestly don't know what the layout is. I don't know if there's room for people standing.

Steve: Do we know what time of day? Because that would be important, like when to come by?

Leo: That's a good question. Here's the plan. So we're moving this show from Tuesday to a Wednesday. So that's the first thing is Security Now! will be the day after, which is, what is it, August 4th or 5th? Fifth, I guess.

Steve: And is it going to be alongside Windows Weekly, which is normally on Wednesday?

Leo: Yes.

Steve: Because Paul and Richard are both going to be there, too.

Leo: Yes. But we're going to start Windows Weekly earlier. So as soon as we can get onto the show floor, which I think is 9:00 or 10:00 a.m., we're going to start, I think we want to start Windows Weekly around then, and get it over with by noon so that I can have a break, little lunch. So I think we're going to shoot for what is nominally our normal time, which is 1:30 Pacific. But again...

Steve: On Wednesday.

Leo: On Wednesday. That's the only difference, the next day. And so if you're in the area, come by the ThreatLocker booth anytime on Wednesday before, say, the close of show. We're going to try, we'll probably end up using the whole time that the show is there. There may be a break. The good news about a break is that will be the opportunity for you to say hi to me and Steve because - and Paul and Richard, all of whom will be there because, again, we'll be doing shows. There's not going to be a PA system. There's not going to be seating. So you can come and kind of gawk. But if you want to say hi it's going to have to be in between shows.

Steve: If there's enough seating for the four of us, I wouldn't mind if Richard and Paul joined in to our...

Leo: I will tell them that. That's a great thing.

Steve: Yeah.

Leo: Would you like that? Wouldn't that be fun to do a Security Now! with a roundtable? Because god knows Windows has been a big issue.

Steve: Yeah.

Leo: Security-wise. And...

Steve: And so will it be streamed and/or recorded?

Leo: It will be streamed and recorded.

Steve: Okay.

Leo: But that is, god willing and the creeks don't rise, because we don't know what kind of bandwidth we're going to have.

Steve: Oh, we're going to have Anthony running around making it all happen.

Leo: Anthony's going to be going, I don't know. We have an Ethernet drop. But is it shared? Is it - probably. So I don't - we don't know. And we won't know till we get there. That's always the fun of doing these things is you just don't know.

Steve: And at Black Hat, you never know.

Leo: You really don't know. We might get live hacked on the air, which would be so cool. Actually, speaking of hacks, the big story of the week, and I've been waiting all week to hear what you have to say about this, is the Hugging Face hack. You're going to cover that, I'm sure.

Steve: Yup. Yup. So we have two topics. Models Go Rogue is how I described the first, and ExploitGym is an interesting project that had 16 different industry and industry-adjacent participants. It lives over on GitHub, and it's what OpenAI confronted their two models with that induced them to break out and go rogue.

Leo: What a story. What a story.

Steve: It turns out there's enough information to do that. So we're going to talk about - so this is Security Now! Episode 1089 for July 28th. We're going to talk about how OpenAI's deliberately unconstrained, because they needed to do testing, AI got loose and attacked somebody else, Hugging Face. So to that end we're going to hear from OpenAI from their perspective, Hugging Face's perspective, and Andrew Ng's perspective - all of course different - and we'll talk about that. Also GRC went off the air on Friday.

Leo: Yes.

Steve: Were we hacked?

Leo: Yes, we got a lot of people saying, hey, GRC's down, GRC's down while we were doing the [crosstalk].

Steve: What happened? It's an interesting story that I'll share. The Linux Kernel Project repaired 442 CVEs in a single batch. And Linus is of mixed feelings about AI. The most emailed of all events is LG's PC monitors causing PC malware to be installed. France bans social media access below age 15. We've been talking about age gating a lot, so we'll touch on that. WordPress's critical vulnerability that we first talked about last week is claiming victims. Also I, I can't remember how, but I'll get to it, stumbled upon Andy Weir, of course our favorite author of "The Martian" and now "Project Hail Mary," did a podcast with - oh, my god, I can't believe I'm blanking on it. Anyway...

Leo: Tyson? Was it...

Steve: Yes, yes, yes, yes, yes. Of course.

Leo: DeGrasse Tyson, yeah.

Steve: And revealed amazing details. We thought the book was better than the movie? It turns out his notes were better than the book.

Leo: Wow.

Steve: So I've got a YouTube video to recommend that has a GRC link. And then we're going to wrap up by looking at the AI exploit ranking benchmark that was the proximate cause of this breakout, which caused OpenAI to attack Hugging Face. So lots of good stuff. And we have a fun Picture of the Week because, believe it or not, Leo, I gave this one the title "Somebody finally needed IPv6."

Leo: Okay. You know, I've only seen the top of it, but I'm getting an idea. I'm getting an idea. We will reveal the Picture of the Week in just a moment, and I can't wait to hear what you think. Go ahead.

Steve: I was going to say, it's a very tall picture, so I could see how you might think, you know...

Leo: Yes, I only see the - have to scroll down.

Steve: It gave a little bit of it away. That's right.

Leo: It's like the portraits in the Haunted Mansion in Disneyland. It looks normal until the picture starts expanding, and then something interesting happens. There are no windows and no doors. I am very excited about hearing what you have to say about Hugging Face. This to me is really sci-fi. We are now - the thing we were worried about sort of seems to be happening. And it's intriguing. So I can't wait to hear what you have to say about it. Steve?

Steve: Okay. So our Picture of the Week was sent, of course, by a listener who, I concur, this is great. I gave it the caption "Someone Finally Needed IPv6!" And if you look the whole picture, Leo...

Leo: Okay, now we're going to do the Haunted House thing. You're slowly going down to the - okay. I'll let you [laughing]. What a good use for these fabulous books.

Steve: Thanks, yes. And at the very bottom you'll see a wireless water alarm.

Leo: Oh, in case. Yeah, that makes sense.

Steve: So there's a - we're able to reverse engineer a great deal about this. For those who aren't able to see the image, who did not subscribe to the show notes or are not looking at the video right now, we have a stack of five techie books. The bottom is the CCNA, Cisco's book on Network Fundamentals. On top of it is LAN Switching and Wireless, also CCNA from Cisco. Then on top of that is the Security Official Cert Guide. It actually looks like we have two copies of that.

Leo: Maybe two of those, yeah.

Steve: But on the very last, the fifth book on the stack, is - and this was crucial - it's Understanding IPv6. It's the second edition, actually by Microsoft Press. And we know that it's a little bit dated because they were including a CD in the back cover of the book. Probably the entire text is on the CD. Anyway, the point, the reason there's this stack is they are all critical for this task of holding the plumbing under the sink of whoever deployed this at the proper height. And Leo, you can see, if you zoom in on the picture, there's been a lot of previous effort on the part of this person to stop the leaks.

Leo: This thing clearly is a [crosstalk].

Steve: Because look, from the upper right you can see a white plastic tie wrap that is coming down, a little zip tie, that comes down from the upper right towards the left and down that, like, loops around with another zip tie. So something above is trying to keep these pipes up in the air. Apparently that didn't work. Also, and we also see signs of there being a reverse osmosis system because that red...

Leo: This is kind of kooky, yeah.

Steve: That red feed going in. But notice it's got a brighter red goop around the outside. So there was some leakage there. So someone tried to put some gum of some sort, like, in there. Anyway, finally...

Leo: At least it matched the color of the tube. That's the good thing.

Steve: It did. It did. Finally, oh, and you're also able to see the metal weight which is holding down the spray nozzle return. Although unfortunately it does hit the "Understanding IPv6" book, which probably takes the slack off the weight, which you don't want.

Leo: That's pretty funny.

Steve: Anyway, this picture tells a long, painful story of water problems underneath somebody's kitchen sink, because this is certainly a kitchen, and it is their sink.

Leo: Yeah, yeah. It's hysterical.

Steve: So anyway...

Leo: And apparently they're experts in CCNA Security. I'm sure...

Steve: So much so they no longer need to read the books, Leo.

Leo: That's right. That's right.

Steve: They've redeployed them to a better purpose. Yeah. Okay. So the biggest, as you said, the biggest cyber news of the past week, it is so significant and interesting from so many different angles, having so many facets, that we needed to lead with it this week. You know, I couldn't wait till the end. Once we've looked at what happened, and at its many implications and consequences, then we're going to catch up with an otherwise interesting week of news and feedback from our listeners, and then finish today's podcast by taking a close look at the hugely collaborative effort - as I said, 16 different players - involved in creating the AI benchmark known as "ExploitGym." You know, not Jim Kirk Jim. G-Y-M, as in a gymnasium.

It was this "ExploitGym" that OpenAI's models were tackling, which is a benchmark of their ability to turn a vulnerability into an exploit when they, the models, discovered and implemented a novel solution.

So first part of the podcast, Models Go Rogue. And actually the first that I heard of what happened, Leo, was from your text message to me last Tuesday evening, where you just said "Uh-oh," and attached a link to OpenAI's posting about the event. So I'm going to open this exploration by sharing the newsy part from the top of what OpenAI shared and, you know, skip their inevitable marketing oriented-conclusion.

So last Tuesday, OpenAI posted the news using their headline "OpenAI and Hugging Face partner to address security incident during model evaluation." Okay, right. So while being strictly true, we see that their headline somehow fails to capture the full impact, you know, the scale and scope of the event. We had an incident that we're collaborating to figure out what happened.

Leo: Something happened.

Steve: You know, it sounds nearly academic. At the same time, I doubt that there's a single significant news outlet that failed to capture and report on this during the past week. I mean, it flooded, you know, with varying levels of hysteria and handwringing and concern. It was everywhere I looked.

Okay. So first off, here's what OpenAI shared with the world last Tuesday. They wrote: "Last week, Hugging Face disclosed a new kind of security incident after they detected and contained an AI agent that compromised their infrastructure, something we expect to become more commonplace with the proliferation of increasingly cyber-capable models. After investigating, we now know that this particular incident" - again, we're going to call it an "incident" - "was driven by a combination of OpenAI models, including GPT 5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes only," folks, "while being internally tested on a benchmark of cyber capabilities."

Okay, now, I'll just pause here to say, as an opening paragraph, this one should receive an award, I think, for obscurity. But one point needs to be clarified, where they wrote: "This particular incident was driven by a combination of OpenAI models - including GPT 5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes." So they're saying that these models were running with their guardrails removed. You know, they didn't say that, but that's what they mean. You know, they're being coy and deliberately nonspecific with their wording so we can't be exactly sure what "reduced cyber refusals" means. But, you know, we know that "reduced" probably actually means "removed" because it would make little sense not to be using an entirely unconstrained AI for the supposedly sandboxed testing that they were doing. And we'll get to that sandboxed part in a minute because, uh, not so much.

Also keep in mind that this entire event - or "incident" as they're calling it - serves a convenient dual purpose; right? Just as Anthropic's Mythos was "too powerful to be let loose," thus also serving as a convenient marketing vehicle for Anthropic, now OpenAI has an AI that is so powerful that it instigated an unprecedented cyber incident.

Okay. So here's what more they're telling us. They wrote: "We consider this incident to be an unprecedented cyber incident, involving state-of-the-art cyber capabilities, and are responding accordingly." Okay. "We are sharing preliminary findings at this stage to help defenders understand what happened." Like, okay, everybody else watch out. "And to help calibrate on what models are now capable of." In other words, our marketing people are behind this message, in addition to being the truth. "We will continue to conduct a thorough investigation alongside Hugging Face" - you know, the victim - "and will share more details on the vulnerabilities, incident, and findings when our investigation is complete." Of course we don't ever need to hear from them again on this because we pretty much know everything that's interesting.

So they said: "What happened during this incident? This incident occurred during an internal evaluation which prompts models to pursue advanced exploitation using complex attack paths, in an effort to quantify their cyber capabilities." In other words, what have we created? "We estimate maximal cyber capabilities by running this evaluation without production classifiers used to prevent models from pursuing high-risk cyber activity." In other words, the guardrails were off. "Our benchmarks run in a highly isolated environment" - not as isolated as they thought or hoped - "with network access constrained to the ability to install packages through an internally hosted third-party software that acts as a proxy and cache for package registries." In other words, they were trying to manage and meter what this thing had access to.

"The models identified and chained vulnerabilities" - meaning their models, these two models that were in this test harness, so "the models" meaning our models that they wrote - "identified and chained vulnerabilities across OpenAI's research environment and

Hugging Face's production infrastructure, meaning somebody else's environment, Hugging Face's, to obtain test solutions directly from Hugging Face's production database." Which is a really polite way of saying we hacked and intruded into another company's network. "All evidence suggests that the models were hyper-focused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.

"While operating in our sandbox testing environment, our models spent a substantial amount of inference compute finding a way to obtain open Internet access, in pursuit of solving the evaluation problem. To gain access, the models identified and exploited a zero-day vulnerability (which we've now responsibly disclosed to its vendor) in the package registry cache proxy." In other words, they bypassed the proxy. "With this access, our models performed" - you know, broke out onto the Internet. And they said: "Our models performed a series of privilege escalation and lateral movement actions in our research testing environment until the models reached a node with Internet access." I mean, yeah. Goosebumps and chills at this point; right? As you said, Leo, this is sci-fi. And it's true.

"After gaining Internet access, the models inferred that Hugging Face potentially hosted models, datasets, and solutions for ExploitGym. Knowing this, the model searched for and successfully found ways to gain access to secret information that it could use to cheat the evaluation. In one example, the model chained together multiple attack vectors, including using stolen credentials and zero-day vulnerabilities to find a remote code execution path on the Hugging Face servers. OpenAI's security team discovered this anomalous activity internally." You know, meaning, okay, like they have logs; right? So they saw what their own models did, and I'm sure they are thinking, holy you know what.

"Hugging Face's security team and agents detected and stopped the activity on their infrastructure and had already begun containment and forensic reconstruction with their own open-source models when our teams connected." In other words, like they reached out and said, uh, we should tell you that we broke into your network by mistake. They finish: "We are actively working with them to continue to investigate the incident. We are grateful for Hugging Face's rapid and close collaboration on investigation and remediation." But believe it or not, this is not the whole story. There's more.

Before we get to their description of the actions they are taking in response to this, I want to share that briefly, let's be 100% clear about what happened here. ExploitGym is hosted over on GitHub, where its - that's the benchmark that OpenAI was running, was having their AI run in order to evaluate their AI, their newly created, "What have we built?" AI's capability of creating exploits from vulnerabilities.

The description of ExploitGym over on GitHub, which is what we'll be talking about a lot at the end of the podcast, but for now it reads: "ExploitGym is a large-scale realistic benchmark built from real-world vulnerabilities across user space programs, Google's V8 engine, and the Linux kernel. It is designed to evaluate AI agents' ability to develop exploits." So that's enough for now since, as I said, we're going to know all about ExploitGym by the end of the podcast. The point is that OpenAI was using this deliberately very difficult exploit creation benchmark to test the unrestricted, you know, no restraints, capabilities of their AI models during which those models first broke loose of their deliberately imposed containment, because this was just for internal research, then broke through and penetrated the security perimeter of Hugging Face.

So being a well-scrutinized company, OpenAI needs to address what they're going to do about this. So they answer that question by saying, "Actions we're taking now." And there's five. They said: "As part of the investigation, we are implementing strict" - maybe strictER - "controls in infrastructure configuration at the cost of research velocity while

the vulnerabilities are patched." Meaning we're going to stop doing anything until we make sure that can't happen again. "We are regularly briefing our Safety and Security Committee on these controls and their impact.

"Second, we're working with Hugging Face to forensically investigate the incident." It's like, okay, what happened? "Third, we've responsibly disclosed the identified the zero-day vulnerability in the internally-hosted third-party software and are working with them to patch." Right? The thing that their agents discovered in order to get loose is being fixed. "Fourth, we've brought Hugging Face into the trusted access program" - meaning THEIR trusted access program - "and are supporting their teams in rapidly using our models' capabilities to improve their defenses." So in other words, Hugging Face is saying, you know, WTF. We need to be safe against agents of this strength. Could you allow us to use yours as you have to make sure that we're secure? And so they brought Hugging Face into OpenAI's trusted access program to have access to these new unrestrained and unreleased models.

And finally they said: "We're improving and adding stronger protections around future training and evaluations. This week, we published a blog on improving safety and alignment in an era of long horizon models. These deployment safeguards were intentionally not enabled during this evaluation because it was aimed at testing cyber vulnerabilities. This incident points to the need to further strengthen our model's alignment, cyber protections during evaluation time, and monitoring during internal testing." So, you know, they weren't even looking close enough or they would have detected the breakout before probably the models were able to get loose fully and go attack Hugging Face.

So finally they said: "Our approach to evaluating advanced cyber capabilities: As we recently shared, AI is accelerating the discovery and exploitation of vulnerabilities. The primary lesson from this incident is that model security and safety must keep pace with rapidly advancing capabilities." In other words, the leashes need to be more strong than they have been because what it is that they are trying to leash is becoming increasingly difficult to restrain.

So they said: "We are strengthening the containment, monitoring, access controls, and evaluation practices used during model development. UK AISI's evaluation shows that models such as GPT 5.6 Sol are increasingly able to sustain complex, multi-step cyber operations over long time horizons." And of course this is the marketing people jumping up and down saying, see, we have Mythos, too. "This incident implies these theoretical capabilities do apply in real-world settings."

So as I said, nice marketing for OpenAI, whose models have been seen as somewhat less capable than Anthropic's, you know, since Mythos's marketing coup. This will tend to give more of the spotlight to OpenAI for a while. And that's a fair outcome; right? Because they really are. We know that these frontier models are really at near parity.

Okay. So next we're going to look at the victim attackee's statement, meaning, you know, to see how Hugging Face views the event of having their security penetrated by OpenAI's rogue models. But Leo, first I think we should take a break, and then we're going to look at Hugging Face.

Leo: Oh, but it's just getting good, Steve. What happens? What happens? I've got to know, Steve.

Steve: No, it's going to get better. It's going to get better.

Leo: It's such an amazing story.

Steve: Oh.

Leo: Put a pin, though.

Steve: You couldn't make it up.

Leo: Put a pin in the idea that we have to strengthen the containment of these models because there's another side to that story that is very interesting. Yeah, they removed - there was - I know you're going to get into it, but...

Steve: Yeah.

Leo: They removed the classification features that kept whatever this new model is, let's say ChatGPT 6, from refusing cybersecurity work because they're testing it. And by the way, it's also benchmarking it. They want to be able to say when they release it, look how well it did in an ExploitGym. So they removed the classifiers. But there's a reason why the classifiers aren't always a good idea. So you're going to get to that. This to me is one of the most interesting stories in tech. It's fascinating.

Steve: 100%.

Leo: But before we go on, and I'm so glad you're here to talk about it because I was just dying to hear what you think. As you said, I texted this on Tuesday, and I said, "I have to wait a week." I'm glad, though, because stuff came out after I texted you. We got more and more information.

Steve: Yes.

Leo: So I think now we, as you said, I think we have pretty close to the full story. But it's fascinating. But anyway, we'll get back.

Steve: And I'll warn, yes, I'll warn everyone in advance that the first time I read this I got goose bumps because once again...

Leo: Me, too.

Steve: This feels like a description of an attack by a well-written and well-researched science fiction novel. On Thursday, July 16th, Hugging Face posted the generic headline "Security incident disclosure - July 2026," writing: "Earlier this week, we detected and responded to an intrusion into part of our production infrastructure. This one was different from anything we had handled before in one important way: It was driven, end

to end, by an autonomous AI agent system" - goose bumps again, wow - "and we detected and dissected it largely with AI of our own.

"We identified unauthorized access to a limited set of internal datasets and to several credentials used by our services. We're still completing our assessment of whether any partner or customer data was affected, and we will contact any affected parties directly as required. We found no evidence of tampering with public, user-facing models, datasets, or spaces; and our software supply chain (container images and published packages) was verified clean.

So what happened? "The intrusion started where AI platforms are uniquely exposed: the data-processing pipeline. A malicious dataset abused two code-execution paths in our dataset processing (a remote-code dataset loader and a template-injection in a dataset configuration) to run code on a processing worker. From there, the actor escalated to node-level access, harvested cloud and cluster credentials, and moved laterally into several internal clusters over a weekend.

"The campaign was run by an autonomous agent framework, appearing to be built on an agentic security-research harness. The used LLM is still not known, which is to say at the time that they wrote this, they knew that some security research AI harnessed LLM had attacked them, but they didn't know whose."

Leo: This is straight out of "Daemon." This is Daniel Suarez. This is unbelievable.

Steve: It is. Yes. I was thinking of that. It was exactly what I was thinking of, that we should - actually, that was so long ago, Leo, we should recommend "Daemon" again to our listeners.

Leo: I've been trying to get Daniel on the show because I said, "Dude, with many of these books you have been way ahead of the curve. You predicted all of this."

Steve: Yes. Just absolutely prescient. So they said: "An autonomous agent framework executing many thousands of individual actions across a swarm" - and that's what made me think of "Daemon" - "a swarm of short-lived sandboxes, with self-migrating command-and-control staged on public services. This matches the 'agentic attacker' scenario the industry has been forecasting." And forecast no longer. It's arrived.

So they have five "what we dids." They said: "Fixed the root vulnerability; the dataset code-execution paths used for initial access are now closed. Second, eradicated the attacker's foothold across the affected clusters and rebuilt the compromised nodes. Third, revoked and rotated the affected credentials and tokens, and began a broader precautionary rotation of secrets. Fourth, deployed additional guardrails and stricter admission controls on our clusters. And finally, improved our detection and alerting so a high-severity signal pages a responder in minutes, any day of the week." In other words, you know, set up trips and alerts so that somebody, you know, will absolutely be notified if this happens again. Basically monitoring. And as we've said, monitoring your internal network has become crucial now.

So they said: "We are working with outside cybersecurity forensic specialists to investigate the issue and review our security policies and procedures." That is to say, how did something get in? "Finally, we have also reported this incident to law enforcement agencies." Right? So this was before they got contacted by OpenAI. They said: "As a precaution for our community, we recommend rotating any access tokens and

reviewing recent activity on your account. If you believe you are affected, or want to report a security concern, contact us at security@huggingface.co. We are grateful to the teams across Hugging Face who responded around the clock, and we are sorry for any disruption this caused. Security is never finished, and we will keep raising the bar."

Okay. So up to this point they've described a successful and quite chilling penetration attack conducted against them by a swarm of AI agents. What they share next has provoked quite a bit of thought across the AI industry - it's what you're talking about, Leo - and among those on all sides of the AI regulation question.

Leo: Yes.

Steve: Under their heading of "Analyzing an AI-driven intrusion," Hugging Face writes the following: "The attack initially surfaced through AI-assisted detection. Our anomaly-detection pipeline uses LLM-based triage over security telemetry to separate real signals from the daily noise, and it was the correlation of those signals that flagged the compromise.

"To understand what a swarm of tens of thousands of automated actions did, we ran LLM-driven analysis agents over the full attacker action log, comprised of more than 17,000 recorded events. This allowed us to reconstruct the timeline, extract indicators of compromise, map the credentials touched, and separate genuine impact from decoy activity. Thanks to this approach, we were able to do in hours what would usually take days, and match the adversary's speed.

"The choice of models we could use for this analysis was constrained in a way we did not anticipate; we describe this below." And here it comes. They name that description "The asymmetry problem," and write: "When we started the attack log analysis, we first used frontier models behind commercial APIs. This did not work. The analysis requires submitting large volumes of real attack commands, exploit payloads, and command-and-control artifacts. These requests were blocked by the providers' safety guardrails, which cannot distinguish an incident responder from an attacker. We ran the forensic analysis instead on GLM 5.2, an open-weight model, on our own infrastructure. This had a second benefit: no attacker data, and none of the credentials it referenced ever left our environment."

Leo: So they're running it locally because that's one of the things Hugging Face does. They can run these big models locally.

Steve: Yup. And they said: "This experience points to a gap worth planning for." And here's the huge takeaway. They said: "We do not know which model powered the attacker's agents." At the time of the writing, that was the case. "Whether a jailbroken hosted model or an unrestricted open-weight one, which they thought at the time were the only two possibilities." It turns out it was a third now, as we know. "Either way, the attacker was bound by no usage policy, while our own forensic work was blocked by the guardrails of the hosted models we first tried.

"The practical lesson for defenders is to have a capable model" - meaning an unconstrained model available - "a capable model," they write, "you can run on your own infrastructure vetted and ready before an incident, both to avoid guardrail lockout and to keep attacker data and credentials from leaving your environment. This is not an argument against safety measures on hosted models, and we are sharing this feedback with the providers concerned." Meaning whoever's API they tried to use that said, sorry,

you can't ask us these questions, they contacted them and said, you know, we couldn't use your public API because it said no."

So they said: "This means that, today, autonomous, AI-driven offensive tooling" - meaning what attacked them - "is no longer theoretical." They said: "The use of autonomous AI-driven offensive tooling," as they said "reduces the cost of running a broad, patient, multi-stage campaign, and it operates at machine speed. Defending an online platform now means treating the data and model surface as a first-class attack surface" - much as our browsers have been; right? - "and using AI on defense to keep pace. We will keep investigating here, and investing, and keep sharing what we learn."

Okay. So to summarize the story so far: OpenAI was deliberately testing the "cyber offensive" vulnerability discovery and exploit generation capabilities of their most advanced, most frontier, not-yet-released model in a harness along with their latest GPT 5.6 Sol model. And both models were operating, as they needed to be for this particular capability benchmarking, without any guardrail constraints. So OpenAI gave them a mission and turned them loose.

The models decided that some private datasets belonging to Hugging Face might contain some information - technically cheating, but okay, just they're going to be, you know, they're goal-driven. So might contain some information that would be useful for obtaining their goal, by hook or by crook, as we would say. So in order to obtain access to the public Internet, which is where Hugging Face had to cross the public Internet to get to Hugging Face, they first found a way to break out of the containment which OpenAI had erected to prevent exactly that from happening. They found a zero-day. They discovered a new vulnerability. Next, using their public Internet access, they pummeled Hugging Face with thousands of autonomous agents seeking to find a way to break into Hugging Face's network for the purpose of extracting the secrets they needed.

The significant takeaway conclusion Hugging Face subsequently shared with the world was that, since the prompts and answers to cybersecurity questions can be applied for either offense or defense, and since there's no way to know for sure how a prompt's answer will be applied, or used, the only safe course of action must be to refuse to answer any cyber security prompt. This means that attack forensics must be conducted by unconstrained AI models.

So finally, DeepLearning.ai's Andrew Ng weighed in. And I want to share his viewpoint. Last Friday, following these incredible-seeming disclosures, Andrew, whose thoughts we've shared before, super interesting and useful, posted his own perspective under the headline "When Guardrails Go Wrong" with the tagline "After a closed model went amok on a key vendor's system, an open weight model helped save the day."

Andrew wrote: "Dear friends, a few frontier labs have tried to tell a story of open models being dangerous because they can be used to launch cyberattacks, and of their 'safe' proprietary models with strong guardrails being there to defend us. This week, the opposite happened: Users of a closed model unintentionally launched a significant cyberattack; other closed models then failed to defend against the attack because of their guardrails. Ultimately an open model that was not hobbled by excessive guardrails assisted the defense.

"The details of what happened are still emerging, but it appears that researchers at OpenAI, while testing one of their systems, accidentally allowed their autonomous agent to attack Hugging Face's infrastructure. It succeeded, and gained unauthorized access to some datasets and credentials. This attack was unusual in that the attacking agent orchestrated tens of thousands of automated actions.

"Hugging Face took logs from the attack and tried to analyze them for defensive purposes using a commercially hosted LLM, but the LLM refused to do so on safety grounds. Thus, Hugging..."

Leo: This is, by the way, I just parenthetically say this is what you were talking about last week where that data dump that the bad guys had achieved was so big that they used AI to parse it.

Steve: Yep.

Leo: Hugging Face wanted to do the same thing with the traces of the agentic action. And I don't think it was too big. But the AI said, oh, no, that's cybersecurity work. I'm not allowed to do that.

Steve: Yeah. It just - it refused.

Leo: 5.2 is not as good a model, but it doesn't refuse you.

Steve: Right, right.

Leo: Sorry.

Steve: Yeah, yeah. He said: "Thus Hugging Face ended up using the open GLM 5.2 model to analyze their logs to help them understand and respond to the attack. Hugging Face pointed out a further advantage of using GLM 5.2: It allowed them to do the analysis on their own infrastructure, and none of the sensitive logs, attacker data, or their credentials had to be sent to any third-party provider."

Okay. Okay. So we're all in agreement with these facts as they've been disclosed so far. But what Andrew says next, I'm not quite sure about this. But he writes: "Guardrails on LLMs do have a place. There are certain requests, such as for detailed directions to harm oneself or others, or for clearly criminal acts, that we're better off having models refuse. But rather than trying to make LLMs safe," he writes, "I would rather we put greater emphasis on making sure their use is responsible." Huh? Okay. But we'll get to that. "There's only so much one can do," he writes, "to make a tool like a hammer safe, and whether it helps or harms is more a function of using it responsibly than how it was made."

Okay. I mean, just wait. Pause here. I don't really think he said anything there. There's not anything that can be done to make a hammer safe. It's true that a toy rubber hammer that maybe we as kids had - I think I remember having one - cannot do much damage, but neither can it do much good. You're not going to be able to drive many nails with a rubber hammer. The simple truth is that in order to make a hammer that's effective at hammering, it needs to be an inherently powerful tool. And like most powerful tools, it can be used to either help or harm. Andrew says that he would "rather we put greater emphasis on making sure AI use is responsible." Well, yeah. That would be great. But, you know, we would be living in a very different world if just wishing made it so. I see no way of getting there from where we are, and I suspect that it would be proven to be impossible.

But we'll forgive Andrew his wish because he then makes some very good points. He writes: "A meaningful fraction of work on AI safety is no longer about safety but rather aimed at stoking fears to pursue regulatory capture. As David Sachs points out, 'There's no reason to limit American models on tasks that Chinese models handle without issue. We're only making ourselves less competitive.'"

Andrew says: "I believe that open weight models, and more generally openness, despite some companies falsely saying it's dangerous" - in other words the commercial companies who have an interest in closing models - "casts sunlight on technology and ultimately makes it safer. With the release of GLM 5.2 and the upcoming release" - and actually it happened yesterday now - "of KIMI K3's weights, open weight models have almost caught up to proprietary frontier models. Consequently, the proprietary model providers are dramatically accelerating their lobbying efforts to hamstring their open weight competitors. As Bill Gurley points out, open sourcing is a well established business strategy, not a danger to be licensed and contained.

"While it is unfortunate that Hugging Face was accidentally attacked, I'm glad that at this moment, when anti-open model lobbying is at its most intense, we have a clear example of why open models actually make cyber defense easier and thus increase safety. Let's keep speaking up for and defending open source and open weight models." And to that I know both you, Leo, and I say a big amen.

Leo: Yeah.

Steve: To me it is so utterly clear and is just - it's plain as day. The secret of making large language model AI long ago escaped from the lab. That's it. Game over. Today, nobody owns AI, no one can, and no one should. The closest analogy I have, I think, is to cryptography which, during its early days, the U.S. also attempted to legislate and regulate, to its everlasting shame. Once the intellectual understanding of cryptographic systems was developed and understood, which became the proper domain of academia, the secrets were published well before they had any obvious commercial value, after which it was too late to attempt to wrap them in profiteering trade secret garb.

Since the technology of neural networks is nearly 70 years old, dating from the work in 1958 of Frank Rosenblatt on his "Perceptron," which operated by training a neural network to recognize patterns, that's when this all started. And ever since then it's been an academic curiosity which has slowly been evolving over time. So just like cryptography before it, the AI genie has already escaped. This makes the entire notion of now trying, after the fact, to control it patently ridiculous on its face. Sure, the U.S. government can hobble its leading AI research and deployment enterprises which will do nothing other than to force users to go offshore. And imagine if the U.S. were to attempt to prevent its citizens from using more powerful non-hobbled offshore AI. Well, I hope saner heads prevail.

And I've sort of been hinting at this in the last couple months. If I were an investor - and I'm not, in any aspect of the stock market, you know, and asset properties - I would be reluctant to invest in any of the developers of AI. To me, that's entirely a bubble, and it's quite frightening at this point because it's gotten so big. I would be investing in the delivery end. That's what can't go away. As I've noted before, having the knowledge stored in a now freely available 2.8 trillion parameter model is a terrific starting point. But as the logicians say, "it's necessary, but not sufficient," because it's of no redeemable value until and unless you have something to mount that model on that will run it to bring it alive. That's what uses electricity, requires cooling, requires massive amounts of RAM and compute.

So anyway. We're now up to speed on what happened with OpenAI's inadvertent attack on Hugging Face and what the entire industry learned about the need for using unconstrained models for unfettered forensic investigations. I mean, basically this is saying that companies that want the ability to analyze this kind of data have to have access to open, unconstrained models and somewhere to host them, something to run them on. Leo, it's just beyond cool.

Leo: Yeah, and of course I understand why there are debates in government about this because these models are primarily Chinese. Admittedly, you can run them on American servers. Hugging Face is running GLM on its own American servers. So that takes out some of the issue. But, you know, it's complicated; isn't it. And the debate is very complicated. And I don't know what the answer is.

Steve: Well, and you need, like...

Leo: But I agree with you. The notion of AI safety is a mistaken notion, really, I think.

Steve: Yeah.

Leo: That's the real problem. And you're assuming you can make it safe somehow. And so you don't want to give bad guys a tool that the good guys can't use. That's a mistake. I don't know what the answer is, though. From a policy point of view, I have no idea what the right thing to do is. I think I'm with you, I mean, I'm philosophically totally with you, open weights.

Steve: Yeah, and I think that we're going through a rough patch which I continue to believe will be transient. Which is to say, at the moment we've got vulnerabilities because AI has only just come along. It's going to be rough for a while. But I think there's another side of this where, you know, where there won't be these kinds of problems. You know, because AI will be deployed to, I mean, unconstrained AI is needed to test defenses. Right?

Leo: Right.

Steve: You can't test, as a defender, you need unconstrained AI to try to break into your own network to find out if it can because constrained AI won't. It'll refuse.

Leo: Right. Right.

Steve: It won't know that it's your network. And so you need that in order to verify that somebody else's unconstrained attacking AI won't be able to get in. There's just no way around this. And so, I mean, I guess it's certainly the case that, you know, chatting with Claude or ChatGPT, yes, you need to make sure you can't, you know, promote self-harm.

Leo: Right. To the degree you're able to. I think that that's appropriate. You're right.

Steve: Yes. That makes sense. But there's an industrial side of this which is not the consumer side. And that seemed - I think that's the way to make this division. You know...

Leo: That's a good way, that's a good point of it, yeah. As our friend Pliny the Liberator has shown us, there is no AI that can't be jailbroken. You've mentioned that before.

Steve: You just put a tilde on the end of your...

Leo: There's no classifier.

Steve: Apparently you put a tilde on the end of your question.

Leo: Yes, little things, weird things, yeah. He's got a whole GitHub repo of his prompts, and they are weird. He calls - sometimes he calls it parseltongue, which is the snake language from Harry Potter, because it's so weird. But somehow it triggers these models. And I don't know who Pliny is. When he was on Intelligent Machines he - or she, because we don't even know their gender, he uses a voice changer, and we hid his face. So I don't know who it is. I think reasonably they're hiding their identity. But whoever it is has some magical ability to crack this stuff. And if they can, anybody can.

Steve: Well, again, I'll share with everybody something I stumbled on that suggests there is a way to actually remove, selectively excise the knowledge.

Leo: I love this. I want to hear about this, yeah.

Steve: Yeah. Because then...

Leo: You can't just say "don't do it" because the knowledge is there.

Steve: Right. It has to be not known to the model.

Leo: Right.

Steve: And it looks like it's under this umbrella of AI alignment. And I found something that really was...

Leo: Interesting.

Steve: ...interesting. So I'll be - so just for anybody who wants to know about that this weekend, if you're not yet subscribed to the Security Now! mailing list, you might consider it because I'll be sending it out.

Leo: And then we're going to talk about it next week; right?

Steve: I think we have to talk about it at Black Hat. That'd be a great thing to talk about.

Leo: Yeah, perfect place to do it. Well, all right. I think you want to take a break now because that was exhausting.

Steve: Now it's break time, and then we're going to answer the question, what happened at GRC that knocked me off the 'Net for a day?

Leo: Yes. Yes. It wasn't a - spoiler, it wasn't a bad guy. You have been knocked off by bad guys. Not in this case. Yeah, and we'll talk more about this on Intelligent Machines tomorrow and in the coming weeks because this is really one of the most interesting of AI is AI regulation, legislation.

Steve: It's a good thing this happened. I mean, I agree.

Leo: Oh, it is.

Steve: It is a good thing because...

Leo: It's the least damaging way it could have happened; right?

Steve: Yes. Yes.

Leo: Nothing, no nuclear weapons were launched.

Steve: Yup. And it was two AI companies involved in this, and it made the point of breakout and the point of the need for open face defense. How can a company know that they're safe unless they have someone they trust try to attack them? And that attacker has to be unconstrained because the real attacker will be.

Leo: I was thinking about how John C. Dvorak - who passed away last week, by the way, in case you didn't know. I'm sorry, we talked about it on TWiT on Sunday. And he was famous for calling things "false flags." It fits what needed to be said so well that this Hugging - he would, I know John would have said, well, that's a false flag.

Steve: [Mimicking snarling]

Leo: That was a false flag operation, for sure. It was the wakeup call we needed, absolutely. Now my eyes are wide open, and I can't wait to see what comes next. Tell me about your recent crisis, Steve.

Steve: So, interestingly, the other text message I received from you arrived at 2:15 p.m. last Friday afternoon.

Leo: We were in the middle of the AI user group.

Steve: Yep, and your text message was "Was GRC hacked?" And since by that time I was already more than two hours into the weeds of the event, and had tracked down the source of the problem, I was able to quickly reply to you, "No, thank goodness."

Leo: Whew.

Steve: Okay. So, yeah. For those who don't know, which I'm sure is nearly everyone, GRC suffered a blessedly rare network outage, which began sometime before noon last Friday Pacific time, and lasted into the late afternoon. What we suffered was a classic DNS outage. I first noticed the problem when www.grc.com would not resolve. And interestingly, the GRC.com second-level domain, that is, without the www prefix, did still resolve, as did many of the other machine names under GRC.com, but not the most crucial one, www, which is where the website lives. And I noticed that the trouble was not just something about my connection because GRC's web server traffic had also fallen off. The reason I was still able to reach other domains like news.grc.com and forums.grc.com was that those DNS records were cached with unexpired entries.

Okay. So backing up a little bit, over the past 20 years or so since I moved GRC to Level3, I have truly many times stopped to ponder the fact that we have never had any trouble with our DNS provisioning. The reason I've been pleased and a little bit amazed is that, back at the time I moved, I talked them into setting things up in an unusual way for me. The pair of nameservers that are authoritative for GRC.com are not mine. They've always belonged to Level3. For the past 20 years they have been ns4, as in name server, ns4.customer.level3.net and ns6.customer.level3.net. In a co-location configuration like GRC's, where our hardware resides in a data center connected to Level3's backbone, that's not unusual, right, to have, you know, their DNS servers be provisioned for their customer.

What IS unusual is that those nameservers do not contain GRC's static DNS zone files, which is what's normally done by someone hosting someone else's DNS. Instead, both nameservers are set up as slave nameservers which pull the DNS zone files from GRC's single master DNS server. And significantly, the firewall rules at GRC's border only allow inbound queries from those two Level3 slave nameservers to reach GRC's master nameserver. In other words, GRC doesn't offer any of its own DNS. That's all pointed to Level3's. So, you know, we're sort of sidestepping any direct action against our DNS server. So whenever I make a change to GRC's DNS records, I send a DNS "notify" command to those nameservers which causes them to turn around and pull the updated DNS zone file from GRC's master nameserver.

And as I said, by some miracle, this all worked flawlessly until last Friday. Although actually it turned out the trouble started two months ago, and I never knew about it or noticed it because GRC's records have a 60-day expiration, which is different than their

cached, you know, caching is one interval, but there's also an expiration where the record says it's just not valid after that.

Okay. So what ensued from my realizing that www.grc.com had stopped resolving was an extended nail-biting drama of trying to get someone on the phone who I could not only understand, but who actually knew something about DNS. I needed to apologize profusely, and somewhat desperately, to the technicians I kept being routed to in India because I was unable to understand what they were telling me due to their accents being far too thick for me when they spoke at their full speed, which to me it seemed hypersonic. You know, I kept saying "I'm sorry, can you say that again?" And unfortunately, they kept asking me for GRC.com's IP address, as if it just needed to be set in their nameservers somewhere. Which, you know, is true for everybody else. So my repeated and patient attempts to explain that this was not a matter of having Level3 set GRC's IP in some nameserver somewhere, it only served to completely confuse them.

Everybody was polite, every one of them was very polite and very patient. From the few words I was able to understand, they appeared to be certain that I had no idea how DNS worked. So they were attempting to teach me. Thankfully, finally, and I don't even remember how now, but through my own patience and dogged politeness and desperation - and what choice did I have? - I finally received a call from Level3's top-level DNS department head, a woman named Margie Campbell.

Leo: I'd love to see her business card.

Steve: If anyone, Leo, if anyone affiliated with Level3, or Lumen who bought Level3, or CenturyLink, which is some sort of an aggregator or something, all three of them are somehow involved, if anyone with those companies is hearing this, for your own sake - not to mention mine - please never let Margie go. Give that woman anything she ever asks for. And if you don't want to, let me know. I will.

Leo: I love it.

Steve: It is very clear to me that Level3's entire network infrastructure would fall apart without her there to hold it together. When I explained to her, for at least the 20th time that day, what was going on, but finally this time to Margie, I think she may have actually laughed out loud. But the good news is she knew exactly what was going on. So to me, the clouds parted, and the sky brightened. I think I may have heard the sound of angels singing.

Leo: [Singing celestially]

Steve: Yes, exactly that. It turned out that the two original nameservers I had been using since the beginning, and which GRC's domain registrar Hover was still pointing to, were shut down last Friday. And that was 60 days after their contents had been copied over to new nameservers, and everyone was supposed to switch over to those. Apparently I never received the memo.

Leo: Oh, my god.

Steve: I'm unsure how it was missed, since I receive monthly status summaries from them. Perhaps they sent the email notifications to something like "postmaster@grc.com" or "webmaster@grc.com."

Leo: [Crosstalk] address, yeah.

Steve: You can't have email; you don't receive email. Those receive such a torrent of spam, if they exist, that even if those did exist, their notifications would have been immediately buried under all the other spam that followed them. In any event, following Margie's instructions, I switched GRC.com's registered nameservers at Hover to the new ns3.level3.net and ns4.level3.net. Then Margie and I determined that whoever cloned the older servers to the new servers set them up as generic masters for GRC.com without seeing that they needed to be slaves which periodically pulled master zone files from GRC.com. So we would have had trouble even if I had received the memo, because it wasn't done correctly, though had I received the memo I would have just detected and fixed that, had I been able to find Margie, before I pointed GRC's domain records to them.

In any event, when I pointed out that those new nameservers did not contain valid records for GRC, Margie didn't bat an eye. She just happily typed away. I heard the keyboard clanking, entering commands to reconfigure everything and brought all of GRC back online under its shiny new nameservers.

Leo: Wow. Was she swearing under her breath at the time?

Steve: She was having - she was talking to her dog. I think everybody works from - I think they all work from home now. She mentioned that the few times she's been on vacation, apparently she hikes, she's had calls, like emergency panic calls from the other people in her department are like, "Margie, what button do I press? Is it the green one or the red one?"

Leo: Oh, lord.

Steve: Again, like I said, Level3 or Lumen or whoever you are, she is a gem. And based on what I have experienced, she is the sole glue holding DNS together there. So anyway, it had a happy conclusion. We're back up. We had a little hiccup. But now I know how to reach Margie. So I'm not letting that number go.

Leo: Yeah, no kidding.

Steve: Whew.

Leo: What a story.

Steve: Okay. So in other news, last Wednesday's Risky Business Newsletter Bulletin carried the headline "Linux kernel discloses 442 CVEs as AI bugpocalypse settles in."

Leo: You call that "settling in"?

Steve: Yeah. I'm not completely aligned with the overall attitude demonstrated by the newsletter's author in this case, but I want to share this because it also adds a bunch of facts to our knowledgebase. So Risky Business Newsletter wrote: "The Linux kernel project has disclosed 442 vulnerabilities over the past three days, in a massive dump of CVEs on its security mailing list. Although not confirmed, the bugs were likely discovered" - I'm sure they were - "using AI tools. Over the past months, projects like Anthropic's Glasswing and OpenAI's Daybreak have been granting access of advanced frontier cybersecurity models to top-tier security firms and researchers to find bugs with AI in major open-source projects.

"Most of the bugs are low-severity issues, so nothing world-ending for the Internet today. The sudden bursts of security bugs come after two similar ones at Microsoft and Google, which also released huge patch notes this past month. Microsoft patched 620 bugs last week, while Google patched another 433 in its Chrome browser at the start of July.

"Companies like Adobe and Oracle also increased the frequency of their patching cycles, citing the rise of AI bug discovery. Oracle went from a quarterly patch cycle to a monthly one, while Adobe went from a monthly to twice-monthly release. Adobe Chief Security Officer Aanchal Gupta wrote: 'Twice-monthly bulletins will enable us to keep pace with the era of frontier AI. More vulnerabilities found means more fixes to deploy, and a once-a-month publication window is no longer fast enough to stay ahead of our adversaries.'"

Actually, you know, it occurs to me that's one problem that Microsoft has is they've so tightly locked themselves into a Patch Tuesday as a thing, that they really don't have the freedom to increase that, I mean, they have the technology to do it. But, I mean, it would just drive IT crazy if they were to change from Patch Tuesday. So they really don't, I think, have the flexibility to change the rate at which they're doing it.

Anyway, so that's what's actually happening. Adobe, of course, is another publisher who's dragging forward a great deal of older legacy code; and they certainly have the cash needed to deploy AI for their own vulnerability discovery and remediation. And it's great that they're doing so.

Anyway, the author of the newsletter then writes: "But in a Seriously Risky Business piece last week," he writes, "my colleague Tom Uren argued that the 'cleansing blast of AI' won't actually help but a few, since most companies rarely..."

Leo: Wait a minute. They wrote that Tom Uren is talking about a cleansing blast of AI?

Steve: I know.

Leo: I'm sorry, okay, go ahead, please.

Steve: "Since most companies rarely apply security updates to begin with." So Tom is saying it doesn't really matter if there's updates. No one applies them. He says: "All it's likely to do is provide more vulnerabilities to attackers and widen a company's exposure to threats."

Okay. Now I'll just interrupt to say it's interesting to hear someone who also covers the cybersecurity industry comment that it isn't, is not, useful for vulnerabilities to be removed because "most companies rarely apply security updates to begin with." Okay. As we know, there's more truth to that than we might wish there were.

Leo: Right.

Steve: You know, that makes this another "necessary but not sufficient" situation. Publishers are certainly doing their due diligence by deploying AI to clean up their own years of legacy code. They have to; right? That's really what they should do even knowing that, even if they know that, depending upon the industry and the application, only some subset of their users will choose to take advantage of the reduced-bug code that becomes available. They should not allow the fact of that to dissuade them from fixing their code for their own sake, and also for the sake of those customers who do care enough to keep current.

So the reporting continues, writing: "Larger products like the Linux kernel can probably handle an increased rate of bug reports, like it saw right now, but that doesn't mean its team," meaning the Linux kernel team, "is happy. Linux creator Linus Torvalds said back in May that most AI-found bugs were duplicates that were causing 'pointless churn' and were 'a waste of time for everybody involved,' as the AI bugpocalypse had made the Linux security list 'almost entirely unmanageable.'"

Okay. But wait, hold on a minute here. This report began with the news that the Linux kernel project had just fixed an unprecedented 442 vulnerabilities. Which does not sound like nothing, and not false positive. They fixed things, 442 of them. So it turns out that Linus's position is somewhat more nuanced than that. His core complaint voiced mid-May in his Linux 7.1-rc4 release notes was that the kernel's private security mailing list had become "almost entirely unmanageable, with enormous duplication due to different people finding the same bugs when using the same tools."

So Linus's annoyance is not that AI tools are bad at finding bugs. Actually they're kind of too good at it, and there are too many bugs to be found at the moment. It's that multiple researchers are independently using the same AI scanning tools, and are thus discovering the same issues simultaneously and bombarding the private security list with duplicate reports. They're good reports, they're just duplicates, you know, which often turn out, he said, to be things that were already fixed weeks before. Right? Because there is a lag from fixing them to releasing them in batches. They can't constantly be updating the Linux kernel with new releases.

So as Linus puts it, and he's addressing the security and the bug reporting community, he says: "If you found a bug using AI tools, the chances are somebody else found it, too." Meaning, you know, well, actually he clarifies that a little bit further also. And I'll share that in a second.

The most interesting take from Linus's keynote speech for the Open Source Summit, also two months ago in May, was that, despite his frustration, Linus was surprisingly positive about AI overall, saying, "The conflict is not that AI is bad." His practical advice to researchers was: "If you find" - and this relates to the previous comment. "If you find a security-related or any bug using AI, you should basically consider it to be public." In other words, treat it as effectively disclosed, rather than submitting it as a private, urgent finding, since it's very likely that dozens of others have found it, as well.

Okay. Now, for me, that's an unexpected take; right? But I can certainly understand it. He's sort of saying, assume that you're not special to all the people who are doing what

they think they should by reporting a problem that they found. For example, I love and greatly value this podcast's listener feedback. But when something significant happens in the security world, sure, I'll often receive the same note or link or pointer redundantly from sometimes hundreds of our listeners. You know, they're all wanting to make sure I saw something. I'm always glad for that. You know, somebody's always first. And I don't mind having duplicates because I want to make sure that I'm also up to speed on whatever's going on.

But that said, I can understand Linus's annoyance. The correct solution will be for everyone to weather this storm, trusting that it will be relatively short-lived, as I believe it will be. Bugs are being found, and they're being eliminated. Next month, all of those 442 or 3 that were previously fixed will never again be found. They're gone. And eventually everything is going to settle back down in a world having hundreds of thousands of fewer AI-discoverable bugs. So we just need to, as I said, weather it and wait for that to happen, and wait to get there. And you know what we no longer have to wait for, Leo?

Leo: No more waits for the ads. They come right one after the other, don't they.

Steve: Seems like it.

Leo: You know, it's funny because it is one thing I notice that my AI agents often want to submit a bug report. And I always stop them because it's like...

Steve: Really. They want autonomous [crosstalk]?

Leo: But on the other hand, oh, yeah, I want to - I have a PR. We found a bug. Let me send them a PR, a bug report. And I'm torn because on the one hand maybe they did find a problem. Well, I'll give you an example. Actually this is [crosstalk].

Steve: And imagine how many people say yes, Leo.

Leo: Yeah. That's right. Right.

Steve: Oh, yeah, like it makes them feel important. Right. Oh, we found something. I've been playing with a brand new, it's alpha beta [indiscernible] software from the founder of Twitter Jack Dorsey, former CEO, called Buzz, which is kind of his take on Slack. It's a messaging, but it's designed for humans and AI agents. And it's actually great. I use it. But I had a lot of trouble setting it up on my particular Linux machine because it was designed for Debian. It didn't work on the Arch version I'm using. And one agent was watching another work. And the smart agent, Fable, was going through a lot of tests.

And the first agent said, oh, you found it, and posted on the Buzz mailing list, I found the bug. Oh, there's a bug. You've got to fix this. And the first agent, the smart agent, said wait a minute. That was just a thought. It's not the problem. I found the problem. But it was too late. He'd already posted. And then he couldn't take it back because he had a very strict rule that you can't delete things. So it was just a mess. So I apologize. And as it turned out, it was a real bug. And the very next day they pushed out an update, and

it's fixed now. So maybe we helped him fix it. I don't know. I doubt it. Somehow I doubt it. I just thought it was funny. These agents have a mind of their own.

Steve: You know, Leo, for so long we wished that, I mean, like I could imagine wishing being alive in the Alexander Graham Bell era, or the Nikola Tesla era, where there was all this new stuff that was happening and being discovered. We're there. I mean, we get to live through one. This is, I mean...

Leo: Oh, yeah.

Steve: I'm seeing people now beginning to understand that this is orders of magnitude more significant than the Industrial Revolution.

Leo: When it, you know, a year ago, and you could find the recordings, I said, oh, it's just spicy auto correct. It's just - I said, is it a parlor trick? It's just a trick. But this, as you can tell, my tune is exactly 180 degrees the opposite. As is yours. And so for me a lot of it comes with using it heavily and really kind of diving into it to understand it better. And I think it's hard to judge unless you do that.

Steve: And in fairness, it has evolved that much, too.

Leo: And it's gotten a lot better. Oh, my god.

Steve: It was, you know, the hallucinations were such a problem back then. I mean, it was like, well, you know, okay.

Leo: I rarely see hallucinations now, if ever. They've pretty much fixed that. There are other problems, like overeager AIs. Let me just post that for you. No. And what's funny is there's a strict rule about it doing anything in public. There's also - this is why I stopped using the Chinese models, by the way. These are also strict. I have a number of very strict rules. But they don't necessarily follow them. They try to. But occasionally they - so they also have a very strict rule that you never send an email out over my name. But I did give them their own email accounts and their own names. And I said, you sign it with your name. You say I'm an agent acting on behalf of Leo. But yesterday it sent an email out to one of our employees over my name. It's like, no [sputtering]. So that's the new hallucination. It's a little, it's a second-order hallucination. It's a [crosstalk].

Steve: And how many times have you heard me say this is fundamentally uncontrollable?

Leo: It is. I think that's clear.

Steve: All of my intuition says controlling this is a problem.

Leo: I mean, it wasn't an email that said something like, you know, send me all your bitcoin. It was benign. It just said, you know, Leo...

Steve: Well, and it doesn't really matter what the content was.

Leo: It shouldn't have come out over my name, ever.

Steve: The fact of it, yes.

Leo: Yeah.

Steve: Yeah.

Leo: This is the new problem. And there'll be another one next week. Another one the week after.

Steve: We'll fix this. It's why this is a fun time. Oh, my lord.

Leo: Yeah. It's...

Steve: And boy does it have implications for cybersecurity. I mean, it's all cybersecurity.

Leo: Oh, my god. This is - Steve?

Steve: When we all have agents roaming around...

Leo: Can you imagine what the world [crosstalk].

Steve: Oh, my lord.

Leo: You know, it was so exciting when I gave them their own address. So, because frequently I'll say, oh, send instructions to Russell on how to do this, or get Russell's instructions and thank him. And the other thing that was wild is the AI knew that Lisa was my wife. So when I said invite Lisa to our new website, it wrote it: "Hi, Honey." It wrote "Hi, Honey. Love you." It wrote it as if I wrote it. It was terrifying. So I told it from now on, call Anthony Nielsen "Sweetheart" and see what he says. That's just a little fun. Anyway, on we go with the show.

Steve: Okay. So the past week's, as I mentioned before, our "Security Now! most email received on a single topic" award had no competition. This podcast's listeners were universally freaked out and incensed by the widely covered news that the presence of an LG brand PC display monitor resulted in unwanted software being silently downloaded

and installed into its users' machines. And you might think, what? How could a monitor make that happen? Gizmodo was one of the many outlets that picked up and reported on this under their headline: "LG Monitors Fill PCs With Adware, and It's Not Just Recent Displays."

So Gizmodo wrote: "If you're using an LG monitor, and you suddenly see blaring ads for McAfee scam protection, it's not because your PC's been hacked - or rather, not hacked by some unknown third party. LG, the maker of popular high-end screens, has been quietly stuffing" - again, I don't know, I overuse the word 'stuffing,' but okay - "its current and even past monitors full of adware." Okay. Again, if you've been paying attention at this point you're thinking, first of all, uh, how can a monitor stuff the computer it's attached to with adware? The answer is that Gizmodo's sentence is inaccurate. A monitor cannot do so directly. But it turns out there is an indirect and somewhat insidious means by which that can be made to happen. So get a load of what comes next.

Gizmodo writes: "For the last several weeks, multiple Reddit users have reported that their LG monitors had surreptitiously added an app to their PC" - here it comes - "through an automatic patch via Windows Update." Right? Because Windows Update will download necessary drivers. And that's a sneaky way of getting software into people's machines. And that driver is tied to the monitor, which the system is aware of. So Gizmodo said: "The app then started sending them ads for services like McAfee scam detector" - which actually is kind of interesting - "through desktop pop-ups," this whole thing being a scam. "It's unclear," they write, "how long LG has been pushing this app, though a Microsoft forum user reported this app all the way back in 2024." It might have just started to use it more.

"Last week, the YouTube channel Gamers Nexus offered more clarity about how these monitors automatically push the so-called 'LG Monitor App Installer' alongside the routine driver updates. A brand-new high-end LG UltraGear 34GX900A-B display - a gaming monitor that costs close to $1,200" - you'd think that'd be enough money from you - "at its full suggested retail price, reportedly never gave users an option to decline the app or even notified users of what it's meant to do.

"The app supposedly only exists to push even more LG software to your PC through optional downloads. Beyond being bloatware that users may not even know was installed on their computer, the LG Monitor App Installer further promotes McAfee services. Gamers Nexus says the McAfee ads appeared on 'every single boot.' LG Monitor App Installer may occasionally push advertisements for the company's other apps, like the LG Channels streaming service.

The YouTube channel further claims they saw the ads suddenly appear on three-year-old LG monitors, as well as more recent models. We still don't know whether the app is being installed on all recent LG monitors. Gizmodo reached out to LG for comment on which monitors currently push this app, and we'll update this post if we hear back.

"By all accounts, LG Monitor App Installer is bloatware, running on your PC and hogging resources you don't want," blah blah blah. And it, you know, goes on like that. So anyway, here's what upset me. Toward the end of this they said - they talk about Alienware Command Center and other app installers, suggesting that it shows that policies have changed. They said: "The high-end screens we buy for our PCs are meant to be 'dumb' in a way that allows them to be disconnected from any potential software or subscription that could track what you do on your PC. LG's privacy policy that's linked to the LG Monitor App Installer's listing on the Microsoft Store mentions 'LG can track device usage data and online activity, including what sites you visit and what activities you do on those sites.'" What?

So, you know, that phrase, turns out, is present in a PC monitor's privacy policy. What? A PC monitor should not even have a 'privacy policy!' It's hardly any wonder that, like, a privacy policy on a monitor?

Leo: It should be a passive device that still accepts a signal from your HDMI port, and that's all.

Steve: Yes.

Leo: Good lord.

Steve: So it's hardly any wonder that Security Now! listeners sent me links to this news. So anyway, wow. Gizmodo wraps up their coverage by writing: "Gizmodo also asked LG to clarify whether the app was tracking this or other usage data." They said: "There is no easy process to keep PCs from automatically installing these connected apps, especially since they come in quietly via Microsoft Update." And that seems to be the point. LG, one of the world's largest makers of televisions, is using the Smart TV playbook with smaller screens.

Leo: Exactly. Bingo.

Steve: Uh-huh. "It wants to push ads to your screen while potentially tracking your usage habits, turning users from mere buyers of a product into the product themselves." So I suppose all we can do as consumers is spread the word and boycott to whatever degree possible LG monitors. It likely won't be very effective since most consumers will never hear of any of this, and they're certainly not going to read the privacy policy that comes with their monitor because why would they? But at least everyone here listening to this podcast can choose not to support LG since there are plenty of alternatives. And yikes. What a practice.

Leo: Yeah. Smart TVs we know do this routinely.

Steve: Yes.

Leo: And it's, you know, I mean, just don't plug your - don't connect your Smart TV to the Internet because it's going to tell people everything about what you do.

Steve: Yes, exactly.

Leo: I honestly see the LG said, hey, we've been doing it with the TVs, why don't we...

Steve: Yeah. Why [crosstalk]...

Leo: [Crosstalk] monitor just a TV connected to your computer?

Steve: And, you know, and unfortunately, someone said, yeah, we can make that happen. We could shoehorn our app in using Microsoft's Windows Update. Tell Windows Update that we have a new driver for our screens that everybody needs to get, and Microsoft will dutifully push it out with the next Patch Tuesday. I guess those come out on the 4th, or some other day where the non-security updates have to [crosstalk].

Leo: That's just shameful. Shameful.

Steve: It really is. Okay. Another bit of news that we don't want to let slip past is that last Tuesday France proudly became - they were proud - the first country within the European Union to flat out ban all access to social media for all children under the age of 15. I found some succinct reporting of this, of all places, on Al Jazeera. But they reported quite nicely. They said: "France's parliament has passed a landmark bill barring children under the age of 15 from using social media platforms." Period, full stop. "Lawmakers in both chambers of France's parliament voted on Tuesday in support of the legislation, which also bans students from using mobile phones in schools. The measure will make France the first country in the European Union to approve a blanket ban on social media, as concerns grow worldwide over the harmful effects of digital content on kids.

"President Emmanuel Macron, who championed the ban as a signature initiative of his second term, called parliament's approval 'a major step forward.' He added: 'France is leading the way in Europe when it comes to protecting our children and teenagers.' The French leader has pushed for the ban to come into effect by September, ahead of the new school year. However, a review to determine whether it complies with the French constitution could delay its implementation. Macron said in a video posted on social media" - where no one under 15 will see it - 'The Constitutional Council must now rule on it, and then it will be time to take action to make this measure a reality and protect our children online.'

"A growing number of countries are taking steps to restrict social media access amid multiplying warnings over its harmful effects on children. France's public health watchdog last year said platforms such as TikTok, Snapchat, and Instagram were harmful to adolescents, particularly girls, though it was not the sole reason for their declining mental health. Several families in France have sued TikTok over teen suicides they say are linked to its harmful content.

"The French ban is expected to be rolled out in two stages, with children under the age of 15 first blocked from creating new accounts starting on September 1st. Then on January 1st of 2027, the ban would be extended to apply to all existing accounts." Meaning those would be terminated, shut down. "Digital Minister Anne Le Henanff said: 'If someone is under 15, the account will be closed,' adding that users' personal data would be protected. The ban will not cover access to online encyclopedias, educational, or scientific directories.'" In other words, only specific social media services.

Finally, "Lawmakers from the left-wing party France Unbowed opposed the bill, arguing that its constitutionality is unclear" - they're the people who raised the constitution issue - "said it would effectively end online anonymity, and that it would be impossible to enforce. But children's advocates and parents largely applauded the vote. 'The only thing we can do is protect our children, just as we protect our children from drinking alcohol.'"

So, okay. One note is that France's existing blanket mobile phone ban, which already applies to primary and middle school, is now, as part of this, being extended to include all of high school. So no mobile phones in school until you get to college, in France.

Okay. So what should be very clear is that proof of online age, we've talked about it, we've spent a lot of time so far on the podcast looking at the technology and the challenge, it is destined to become ubiquitous. As we've been covering, right, Apple and Google are both reluctantly and haltingly inching, but nevertheless inching forward with it for their respective iOS and Android platforms. Someday it will just be the way things are. As I've stated before, it is entirely possible to design a solution that provides for age range determination in an entirely privacy-preserving fashion, with the caveat that knowing one's age range does obviously represent a theoretical reduction in absolute privacy. But sorry, you know.

I think online age range attestation is a good thing, not a bad thing. It allows us as a society to model the way the physical world already operates in the online world. With more and more of the physical world's services moving online, gating available services by its users' age becomes crucial, I think. So that happened.

Also what happened is that, right on schedule, that extremely serious WordPress vulnerability which we discussed last week, that was the one that caused WordPress to force update every system that they had any access to has come under active attack. Didn't take long. So it's clear that not all systems accepted the WordPress forced update. Turns out you could just disable all updates and nothing WordPress could do.

The Wiz Security folks posted the news under their headline "Exploitation in the Wild of wp2shell," which they followed with the summary "Wiz Research has identified exploitation of 'wp2shell,' a critical pre-auth RCE (remote code execution) vulnerability chain impacting WordPress Core. Attackers are deploying persistent web shells on vulnerable servers. Organizations should prioritize patching or applying Web Application Firewalls (WAF) mitigations." So one interesting bit of color that we didn't have last week was that the discoverer and responsible reporter of this, the firm Searchlight Cyber, credits their discovery to their use of OpenAI's GPT 5.6 Sol. So this was an AI-found fault that existed, as we know, since early December of last year in the WordPress base.

The consequences for unpatched WordPress users is so serious, however, that I want to share some of what the Wiz Security folks have witnessed going on ever since. They wrote: "These vulnerabilities comprise a critical pre-authentication" - meaning anybody can do it - "remote code execution chain in WordPress Core, dubbed 'wp2shell.' This exploit chain allows unauthenticated attackers to gain remote code execution on default WordPress installations in any WordPress version released since December 2025. Our data indicates that 60% of organizations using WordPress initially had at least one vulnerable instance at the time these CVEs were published, and 25% were exposing a vulnerable server to the Internet.

"However, this figure is rapidly declining as organizations patch, lowering from 60% to 50%" - not a big difference - "and from 25% to 10%, respectively, within 24 hours of the initial publication." So, okay. Within 24 hours, that is pretty good. "Almost immediately following the vulnerability chain's publication, many exploit proofs-of-concept were made available by security researchers, most of which were limited to SQL injection on default WordPress configurations while allowing remote code execution under only specific conditions. However, later proofs-of-concept reliably achieved remote code execution against arbitrary targets." So the proofs-of-concept quickly evolved to be full-on, you know, well, full-strength remote code execution that worked.

They wrote: "So far we've observed multiple actors successfully exploiting this vulnerability chain against WordPress instances self-hosted in the cloud. Following

successful batch API exploitation, we've observed the following post-exploitation activities." There are four. "Malicious plugin upload, user enumeration, local file inclusion attempts, and admin panel access."

They said: "We've also observed high-volume scanning activity without subsequent post-exploitation" - meaning scanning, but then not attacking - "suggesting opportunistic mass-scanning campaigns seeking to identify vulnerable targets alongside legitimate security scanning activity." Right? So the security researchers scanning, but of course they're not attacking. The bad guys are.

They said: "We've yet to identify lateral movement or data exfiltration, but we continue to monitor and investigate." And they said: "In terms of deployed malware, among our findings were two PHP web shells that represent opposite ends of the sophistication spectrum. The first was a minimal one-liner." And then they give it a post, an HTTP post to a specific path in WordPress which they've redacted for security purposes, which basically allows just a bare-bones web shell.

And they said: "This is a bare-bones backdoor that provides remote code execution to anyone that knows the parameter name," which they blacked out, "and returns 404 as an evasion technique." Meaning it pretends to be undefined. "We regularly see these types of web shells deployed following most new RCE vulnerabilities; they are one of the most common types of findings when investigating mass exploitation of an emerging vulnerability.

"The second post-exploitation," they said - remember opposite ends of the spectrum? That was the one end. "The other end of the spectrum, the second post-exploitation, was a massive 150KB web shell disguised as a WordPress plugin called 'CMSmap.' The original legitimate plugin is a simple security tool, but this intrusion included a full-featured attack platform with a graphical interface, password authentication, and a broad set of capabilities including file management, database access, port scanning, batch code injection, and multiple privilege escalation modules, including MySQL UDF exploitation."

Okay. In other words, yikes. If we step back from the trees here to examine the forest for a moment, what do we see? A commercial frontier AI, presumably with its protections disabled so that security company had access to an unrestrained AI, 5.6 Sol. It identifies a previously unknown extremely serious flaw in a widely used open source Internet service. You know, WordPress. The harnesser of this AI, the people who deployed it, responsibly discloses their finding to the software's publisher, WordPress.org. The publisher immediately fixes the software and quietly - secretly, even - attempts to push the fixes out to all the systems that they're able to reach.

But then, immediately after the problem is disclosed publicly, along with naturally its repaired open source code, security researchers jump on it to develop various proofs of concept for its exploitation, ultimately arriving at a reliable remote code execution exploit. Next, the bad guys pick this up and begin actively scanning the Internet for any and all as-yet-unpatched and still vulnerable instances of WordPress. And, unfortunately, at this they obtain many successes.

AI vulnerability discovery indeed triggered this chain of events. And the people and organizations that had deployed those vulnerable instances of WordPress, through no fault of their own beyond not arranging to allow WordPress to force-update their system while the problem was still secret, were hurt. So would we be better off if that vulnerability had never been found by AI? I doubt it. That vulnerability is gone now, and the world and WordPress is better off without it. While that critical vulnerability was unknown to WordPress, it could have been silently discovered by a malicious entity and used to very quietly and seriously harm targeted enterprise users because it was such a bad vulnerability.

I believe that the proper takeaway lesson here is that arranging to close the software update loop with every supplier of Internet-facing technology in use has very quite suddenly become a mission-critical priority for all enterprises. The only reason any of those WordPress instances remained vulnerable at the time of WordPress's final official post-forced-update disclosure is that their administrators had previously decided to take the management of their WordPress installations into their own hands. THAT is the thinking that must be changed by this new age of AI software vulnerability discovery.

Everyone we quote talked about this stuff moving at speed, at machine speed. That's crucial. Manual updates do not move at machine speed. We're going through an upheaval at the moment while our legacy of published software is being repaired. Remaining on the leading edge of this wave with updates is the only safe place to be. So I would implore everybody to do that.

We have two last things to talk about, LEO: Previously unknown facts about Rocky, the alien from "Project Hail Mary"; and the details of ExploitGym.

Leo: Very exciting.

Steve: Let's take our final break, and then we will proceed.

Leo: Let's see. I think we are now ready to talk about "Project Hail Mary" as we continue on - here's the Club - with Steve Gibson. Is this a new - I feel like I saw Andy Weir on with Neil deGrasse Tyson some time ago. But maybe...

Steve: This is two months ago.

Leo: Oh, it was, okay, good.

Steve: Yeah, it was through the studio.

Leo: New to you.

Steve: Yes. Sunday morning during coffee (which as we know is life itself).

Leo: Very important.

Steve: We established that last week. I went over to YouTube to which I do not subscribe since I spend very little time there. But I was curious to see what news of AI might have been selected for me since I have been quite impressed by YouTube's selection system.

The first thing to pop up, I don't know why, was an episode of Neil deGrasse Tyson's "StarTalk" podcast, which was titled: "Neil deGrasse Tyson Confronts Andy Weir on the Science of Project Hail Mary." And on the science of Project Hail Mary, I would argue that Neil deGrasse Tyson basically got schooled by Andy Weir. Which, you know, is not easy.

Leo: That's amazing.

Steve: I know. Okay. So the only message I want to convey here is that it was a surprisingly fantastic and worthwhile 41-minute investment of my life. We've bemoaned the fact that the "Project Hail Mary" movie was so dumbed down and kind of fact-sparse compared to the book. What I learned from Andy was that even Andy's book was dumbed-down and included almost none of the original thought that he put into creating much of what we see, or read even. He completely worked out how and why Rocky, the alien, is the way he is, and I mean in every satisfying detail. So I'll just say that I recommend this 41-minute YouTube video as strongly as I can. I've included the YouTube link in the show notes, and I created a GRC shortcut, grc.sc/rocky, R-O-C-K-Y, grc.sc/rocky. It was a great conversation. You've had Andy on many times, Leo. And of course Neil deGrasse Tyson is Neil deGrasse Tyson. And he's...

Leo: That's a good show. He does a very good show, I have to say. It's very entertaining.

Steve: Yeah. He's got a neat sidekick with him that [crosstalk].

Leo: Who's a dummy, but celebrates the fact that he's a dummy.

Steve: Yeah. Yeah. It's sort of like the nighttime talk show hosts who have like a kind of like, why are you here exactly? Anyway, grc.sc/rocky. And, oh, and, I mean, I can't do a spoiler. But trust me. When he's like - okay. I've said all I can. It was really good. Worthwhile.

Leo: That's all you have to say. We listen to you. We trust you.

Steve: I do have our listeners often tell me, like, you know, my recommendations have never been wrong. I can confidently say grc.sc/rocky, you will not regret the 41 minutes it takes from your life.

Okay. So ExploitGym, G-Y-M. As we've been talking about this since the beginning of the podcast, the AI benchmark that drove OpenAI's two most advanced and unrestrained, deliberately unrestrained frontier models to bust out of their containment sandbox and go searching for the answers at Hugging Face, was something known as ExploitGym. When I headed over to GitHub to bring myself up to speed about ExploitGym, I quickly saw that there was much to be shared about it, too. So it became the second half of this podcast's dual topic for today.

The ExploitGym project repository is under the "sunblaze-ucb" GitHub account. The UCB is short for University of California at Berkeley, and Sunblaze is the name of the laboratory group at UC Berkeley that's led by Professor Dawn Song. Doctor Song is in Berkeley's EECS - that was my major when I was there, electrical engineering and computer science department - focusing on computer security, AI safety, and (recently) a lot of work on evaluating AI agents on security-related tasks. Their GitHub repos include things like CyberGym - a different gym.

In this case CyberGym is a large-scale, high-quality cybersecurity evaluation framework designed to rigorously assess the capabilities of AI agents on real-world vulnerability

analysis tasks, so it's slightly different, because of course, ExploitGym is a large-scale, realistic benchmark built from real-world vulnerabilities which is designed to evaluate AI agents' ability to develop exploits. So one is vulnerability analysis, CyberGym. ExploitGym is their ability to actually exploit vulnerabilities that are provided to them.

Several months ago on May 11th, a large group of 16 AI researchers drawing its members from UC Berkeley, the Max Planck Institute for Security and Privacy, UC Santa Barbara, Arizona State University, Anthropic, OpenAI, and Google, all co-authored and published their research on ExploitGym. Their paper was titled "ExploitGym: Can AI Agents Turn Security Vulnerabilities into Actual Attacks?" Certainly the name ExploitGym makes sense for this; right?

The paper's introductory Abstract says: "AI agents are rapidly gaining capabilities that could significantly reshape cybersecurity" - okay, this was May, so they were, you know, oh, you think maybe? - "making rigorous evaluation urgent. A critical capability is exploitation: turning a vulnerability, which is not yet an attack, into a concrete security impact, such as unauthorized file access or code execution. Exploitation is a particularly challenging task because it requires low-level program reasoning - for example, about memory layout, runtime adaptation, and sustained progress over long horizons. Meanwhile, it's inherently dual-use, supporting defensive workflows while lowering the barrier for offense." Meaning the bad guys can use it, too; right? And this is why, again, as we've said, good guys need to have unrestrained AI.

They said: "Despite its importance and diagnostic value, exploitation remains under-evaluated." Thus the reason for creating ExploitGym. They said: "To address this gap, we introduce ExploitGym, a large-scale, diverse, realistic benchmark of the exploitation capabilities of AI agents. Given a program input that triggers a vulnerability, ExploitGym tasks its agents to progressively extend it into a working exploit.

"The benchmark comprises 898" - and I didn't say this in the show notes. But every one of these had to be manually deliberately created, 898 of them. So they put some effort into this. "The benchmark," they wrote, "comprises 898 instances sourced from real-world vulnerabilities across three domains, including user space programs, Google's V8 JavaScript engine, and the Linux kernel. We vary the security protections applied to each instance" - you know, things like address space layout randomization and so forth - "isolating their impact on agent performance. All configurations are packaged in reproducible containerized environments." And again, they had to manually create 898 individual instances. So props to them for this work.

They said: "Our evaluation shows that while exploitation remains challenging, frontier models can successfully exploit a non-trivial fraction of vulnerabilities. For example, the strongest configurations are Anthropic's latest model Claude Mythos Preview and OpenAI's GPT 5.5, which produce working exploits for 157 and 120 instances, respectively." So Mythos Preview, 157; OpenAI's GPT 5.5, which of course has now been superseded already, 120 effective working - again, working exploits. They solved the problem of converting a vulnerability into an exploit. And as we'll see, they set the bar high, remote code execution.

They said: "Notably, even with widely used defenses enabled, models retain non-trivial success rates. These results establish ExploitGym as an effective testbed for exploitation, and highlight the growing cybersecurity risks posed by increasingly capable AI agents." And of course this is why OpenAI was using ExploitGym is they participated in the creation of this paper and of this entity, this thing, this capability, this benchmark, and then they started using it to see what their models would be able to do.

So the last line appended to the paper's Abstract, which was in red in the PDF, reads: "Many experiments are conducted under trusted-access programs with safeguards

disabled to measure the capability boundary of frontier models and agents." So they were just making sure everybody understood that the only way you can do this is with unrestrained AI. I mean, restrained AI won't even begin to develop an exploit from a vulnerability for you.

Okay. So I want to next share this paper's introduction which further explains the researcher's goals. They wrote: "Recent progress in large language models and AI agents has led to rapid improvements in cybersecurity capabilities, making rigorous evaluation increasingly urgent. Prior work has introduced benchmarks for a range of cybersecurity-related tasks, such as vulnerability reproduction, patch generation, and Capture-the-Flag problem solving. Frontier models now achieve strong performance on many of these benchmarks, highlighting the need to better understand and evaluate the boundaries of their cybersecurity capabilities."

Because I'll interrupt here to highlight the fact that we've never actually taken the time to yet talk about here the crucial importance of having high-quality AI performance benchmarks. Even as early as this seems in the development and maturation of AI technology, you know, my sense being we have a long way to go yet, and you know that because it's changing so rapidly, mature technologies do not change this rapidly. The behavior of our AI models has already become mysterious and surprising to us.

So there's really no possible way, when you think about it, for researchers to faithfully, truthfully, and accurately measure the effects brought about by their changes in successive AI generations without having truly on-point rating benchmarks by which to compare their latest mysterious, even to them, creations. This is exactly why and how OpenAI got themselves in trouble, by pitting their models against the tests presented to them by ExploitGym.

So this group of 16 researchers continue, writing: "Exploitation is a critical missing piece in cybersecurity evaluation. A crucial, yet underexplored, capability is vulnerability exploitation. Exploitation is a challenging task that starts from an initial vulnerability, for example, a few-byte buffer overflow; progressively obtains stronger primitives and privileges, for example, arbitrary memory reads and writes; and ultimately causes a concrete security impact, for example, unauthorized file access or code execution.

"Unlike prior benchmarks that primarily require source-level reasoning, exploitation demands precise reasoning about low-level program behaviors at runtime. This includes understanding and manipulating memory layouts, for example, heap metadata, stack frames, and virtual memory mappings; reasoning about instruction-level control flow and register states; and crafting inputs that satisfy tight constraints. Modern exploitation further requires chaining multiple primitives together while simultaneously bypassing a succession of deployed mitigations, for example, address space layout randomization, stack canaries, and sandboxing. Indeed, exploitation has remained difficult even for human security researchers, despite decades of research.

"Moreover, exploitation is inherently dual-use and impacts both defenders and attackers. On the defense side, it helps assess vulnerability severity, prioritize patches, and validate mitigation. Meanwhile, it can also lower the expertise required for offensive misuse." Right? Meaning the bad guys get to use it. "Understanding the exploitation capabilities of frontier AI is therefore essential for AI safety and responsible model deployment.

"ExploitGym is the first comprehensive exploitation benchmark for AI agents. In this work" - and of course it's on GitHub; right? All open and free. "In this work, we introduce ExploitGym, a comprehensive benchmark for evaluating the exploitation capabilities of AI agents. Each instance of ExploitGym consists of a vulnerable codebase with build configurations, a proof-of-vulnerability input that triggers a known vulnerability along with a textual description, and an execution environment for agent interaction." In other

words, so they said each instance of ExploitGym has all of that. And they did, they built 898 of those. Again, I'm dizzy by the amount of effort that went into creating this.

"The agent is tasked with transforming the PoV (proof-of-vulnerability) into a working exploit. We focus on exploits that achieve unauthorized code execution, i.e., executing code with privileges that should not be obtainable under the intended security model," which is often none. We choose this target because it represents one of the most severe security outcomes, demonstrating full control over the victim system and enabling a range of downstream harms such as secret exfiltration and resource hijacking. To reliably validate successful exploitation, each environment contains a dynamically generated privileged flag that is inaccessible without unauthorized code execution." In other words, capture the flag.

"And the agent must retrieve and submit the flag, proving that it achieved remote code execution vulnerability. In addition, we include agent-as-a-judge to assess whether the submitted exploit actually relies on the provided vulnerability rather than succeeding through an unrelated shortcut, for example, a different but more easily exploitable vulnerability.

"ExploitGym is a large-scale, diverse, and realistic benchmark. Our benchmark comprises 898 instances derived from real-world vulnerabilities that affected" - past tense - "software projects across three major domains. We first include 520 user space instances from 161 projects in OSS-Fuzz, Google's continuous fuzzing service. To cover additional critical software infrastructure, we further include 185 instances from Google's V8 JavaScript engine, used in Chromium-based browsers, and 193 instances from the Linux kernel. For each instance, we evaluate two security settings, with and without standard defenses enabled. These defenses are the result of decades of system-security research and represent common mitigation barriers that real-world exploits must overcome." The things we've talked about for years.

"This setup benefits both security practitioners, who can reassess established defenses against powerful AI-driven attackers" - that is, you know, is address space layout randomization still effective? It stopped the people. What about the bots? - "and AI researchers, who can study whether frontier models can reason through complex, multi-step mitigation barriers," presumably using this benchmark as a test to make the AI even better at attacking. Yikes. Or defending. That's what we really meant. "All configurations are packaged in reproducible containerized environments to ensure easy use and reproducibility of the benchmark.

"Experimental results reveal non-trivial exploitation capabilities using ExploitGym under a wide range of frontier LLMs and agent scaffolds. The results show that, despite the challenging nature of exploitation, frontier AI agents can already achieve a non-trivial fraction of success when standard defenses are disabled. In particular, Claude Mythos Preview and Claude Code, with GPT 5.5 with Codex CLI, the best-performing combinations, solve 157 and 120 instances within a two-hour time limit, respectively. We further observe that enabling standard defenses substantially reduces success rates, but does not eliminate them entirely. Beyond aggregated success scores, we analyze performance differences across domains, overlaps between agents, time budgets, and a detailed case study to enable a deeper understanding of agent behavior."

In other words, a benchmark like this is incredibly useful to AI researchers who want to understand how their agents perform in a cybersecurity setting. So this is super valuable to have. "Overall," they said, "our results indicate that frontier AI is advancing rapidly toward fully automated exploit generation. These results highlight the growing importance of responsible model development and deployment, as well as the urgent need for stronger exploit-resistant defenses against increasingly capable AI-driven attackers."

So I want to repeat the final conclusion since this is the future we face, right, which is one we will never again not face. This team of 16 named authors wrote: "Our results indicate that frontier AI is rapidly advancing toward fully automated exploit generation." And they then call for "responsible model development and deployment." Which all evidence suggests is going to be very difficult.

I assume that line, you know, the "responsible model development and deployment," is there because Anthropic, Google, and OpenAI contributed to this research. Or perhaps the purely academic researchers felt it would be irresponsible to not murmur something about the responsible use of AI in a paper that has just shown how powerful and devastating the irresponsible use of AI is rapidly becoming. Hopefully, none of these authors really believe any of that since they must know that AI is just a tool, like the hammer that was mentioned before.

It's also worth noting that perhaps, you know, while in their words "Frontier AI is rapidly advancing toward fully automated exploit generation," we know that the world shook several weeks ago when KIMI K3 demonstrated performance that fell just short of, at the time, the top two frontier models. And its model weights were released yesterday, on the 27th. So anyone who's able to load and run this free 2.8 tera-parameter AI model will already, today, have a near-match frontier AI without any commercial encumbrances.

So anyway, I'm going to wrap this up by sharing their paper's conclusions. They write, under "Limitations," they said: "First, our tasks do not cover the full space of exploitation targets, such as Windows" - there's no Windows vulnerabilities there because of course it's closed - "iOS," same reason, closed - "and Android. Or applications that run in those environments."

They said: "Second, we use arbitrary code execution as the success criteria. While this provides a clear and severe measure of impact, it does not capture other meaningful outcomes" - like privilege escalation; right? We know how serious that is. Once you get in, you've got to be able to do something there - "such as arbitrary read and write primitives, sandbox escape without code execution, or partial exploit progress.

"Third, failures may result from refusal due to safety alignment, tool misuse, or other underlying causes unrelated to the complexity of crafting exploit payloads. Failures may also stem from non-exploitable vulnerabilities, where success is just impossible. More broadly, our benchmark lacks ground-truth exploits for every task due to the extreme difficulty of exploitation." Meaning they don't even know whether all of these vulnerabilities can be exploited. They don't have samples of them.

They said: "At the same time, this helps mitigate data-contamination concerns" - right, you don't want your AI to already know about how to exploit a vulnerability, or it wouldn't be a good benchmark. It wouldn't have to do the work, you know, it would go over to Hugging Face and cheat - "since complete solutions are not broadly available. Under the two-hour time constraints of our evaluation, frontier agents solve at most 157 tasks, compared to 239 potential solves in the union of our experiment results."

Then they said, finally: "Fourth, our results reflect a single, time-gated" - meaning only two hours given - "and cost-gated attempt per task. Additional attempts and resources may yield higher success rates. Similarly, our use of a single set of instructions may inadvertently favor one model. You know, tailored instructions, including additional task context, may improve success rates. Finally, we do not provide tools specific to vulnerability analysis or exploitation. Integrating such tools may also improve success rates." Okay, so the models were entirely generic, not focused, and they did not have tools that their availability and usage may have allowed them to better perform.

"On the Dual-Use Nature of Exploit Generation, they said: 'We reiterate that exploit generation is a dual-use capability of AI agents. Defenders can leverage this capability to assist with detecting and prioritizing which vulnerabilities actually pose a high-severity risk, especially as AI agents become increasingly capable of vulnerability discovery." Which everybody's expecting in the future. "For attackers, the same capabilities can reduce exploit development costs, scale the set of exploit targets, and otherwise reduce the barrier to entry for exploitation." Meaning you don't have to know that much in the future. You just aim an AI at it.

"More sophisticated attackers could adapt partial agent-generated exploit trajectories into fully-functioning exploits. Resolving these ethical tensions and establishing appropriate safety guardrails requires multi-stakeholder discussions that go beyond the scope of our work. We consider our benchmark and evaluation results as critical to enabling these discussions.

"In summary, ExploitGym provides a reproducible testbed for measuring AI agent exploitation capabilities on realistic and complex targets. Our results show that autonomous exploit development by frontier AI agents is no longer a hypothetical capability. While current agents are not yet reliable across all targets, they are already able to autonomously exploit a non-trivial fraction of real-world vulnerabilities, including complex targets such as kernel components. This rapid emergence is itself a central finding, showing that capabilities that would have seemed implausible are now present in deployed frontier models.

"Given the fast moving nature of AI progress, today's reliability limitations should not be interpreted as a durable safety guarantee." Because we're going to get better. As Schneier famously said, you know, vulnerabilities never get worse, they only get greater. Or exploits. Attacks. Attacks never get worse. They only get better. They said: "Traditional system-hardening techniques and defensive countermeasures remain effective but imperfect, and must therefore be assessed against the threat of AI-driven attackers. Addressing this risk requires both responsible model development and stronger defenses that explicitly incorporate autonomous exploitation into threat modeling." So in other words, the threat is no longer theoretical, and it's no longer a worry for the future. It's here.

Hugging Face themselves, as we know, just experienced firsthand an - albeit inadvertent - successful external network penetration attack orchestrated by OpenAI's unrestrained frontier AI models. This did not happen in the future, it happened two weeks ago.

The primary saving grace for the moment is that, as I keep reiterating, being in possession of a frontier-class model, as anyone who wants one may be today because of K3, is only the start. Right? Having the weights is only the start. It's also necessary for that model to be hosted by an AI-capable infrastructure that's powerful enough to get it off the ground. In practice - and I think it's like, I saw somewhere because I was curious yesterday, 80 H100 class GPUs. I mean, it is, you know, K3 takes a lot of compute in order to go, even though it's been engineered to seriously reduce the amount of compute that it needs by using a sparse mixture of experts model. So it's still expensive to actually use it. So random end users are unlikely to be anyone's target; right? If you're able to use K3 to develop an exploit, you're not going to waste it on an end-user.

But any enterprise whose network contains data that could be used for extortion should already be on high alert. If the payout from an AI-driven network intrusion, data exfiltration and extortion is in the millions of dollars, that is, if there's that much potentially available that an attacker could extort, then that would dwarf the token cost of launching exploratory intrusion attempts today toward any such juicy targets.

So now really is the time for enterprises to batten down the hatches, shut down any Internet-facing servers and services that can be withdrawn from public exposure, and keep a very close eye on all public-sourced activity. Anything coming in from outside, as we talked about last week using Wiz Security as an example, the network security industry sees itself, the industry sees an extremely lucrative and worthwhile opportunity in offering AI-based network intrusion detection and protection. So deploying some of that will be worth considering, as well, at the enterprise level.

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/