AI Captured the Flag

Description: As expected, UniFi OS devices are under attack. CISA commands federal agencies to update Drupal. Can the largest-ever botnet be killed? Defender endpoint can cut off a PC from the network. Charter Communications big account leak. Chrome moves device-bound session cookies from beta. Anthropic to release Mythos shortly. cURL and Daniel Stenberg. IBM and Red Hat commit to fixing open source with AI. LOTS of terrific listener feedback this week. AI spells the end of a terrific source of training.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1081.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1081-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson's here. Lots to talk about. The end of a very popular hacker competition. Why? AI, that's why. The other shoe drops on two things Steve warned us about last week, two big flaws. And why did Pwn2Own says Steve's accounts have been breached? Well, it was a mistake that Steve will explain. That's all coming up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1081, recorded Tuesday, June 2nd, 2026: "AI Captured the Flag."

It's time for Security Now!, the show where we cover the latest security news, your privacy, how things work in the real world with this guy right here, Mr. Real World himself, Steve Gibson, wearing his AI hat. Steve.

Steve Gibson: Mr. All Too Real. That's right.

Leo: Good to see you. What's up this week on Security Now!?

Steve: So for Episode 1081 for this first podcast of June, there was a really interesting story about an aspect that we've not talked about a lot, which are Capture the Flag competitions. Now, we've talked about Pwn2Own, where research hackers find unknown vulnerabilities in existing software. Capture the Flag is different. These are canned competitions where the designer of the competition inserts something to be discovered which is dependent upon the talent of the hackers and researchers. So we're going to talk about that because someone who's been very involved and has got some real street cred for his ability to find vulnerabilities is bemoaning the death of Capture the Flag brought about by AI. So today's title is "AI Captured the Flag." We've got a bunch of neat...

Leo: I hope you're going to explain what all of that is.

Steve: Yes. We'll explain it all.

Leo: Yeah.

Steve: As expected, UniFi OS devices immediately came under attack. CISA has commanded federal agencies to update their instances of Drupal. Both things we talked about last week. Can the largest botnet ever seen be killed is an interesting question we'll look at. Defender, Microsoft's Defender, the Defender for Endpoint is obtaining the ability to cut off an infected PC from its own network. Charter Communications has had a big account leak. Chrome has adopted device-bound session cookies, moving it from beta. We'll talk about that. The news is that Anthropic is about to release Mythos.

We've also got some interesting commentary from Daniel Stenberg, famously the creator/author/maintainer, you know, chief of cURL. IBM and Red Hat are committing to fixing open source using AI and putting I think it's $5 billion worth of AI cred into the project. We've got, in addition to all that, a ton of terrific listener feedback which will drive some additional commentary. And then we're going to look at AI basically spelling the end of what has turned out to be a terrific source of on-the-ground training for hackers and researchers. And of course a great Picture of the Week. So I think maybe worth staying tuned in.

Leo: Yay. As usual. This is the day you look forward to all week long, if you're in the IT business. You certainly need this to keep up on what's going on. And we thank Steve for doing that. We will get to, of course, the Picture of the Week, I'm sure a humorous take on the world of...

Steve: On humanity.

Leo: ...of humanity, yeah, no kidding.

Steve: So did you guys talk about ownership changes?

Leo: I did. And he said we're absolutely committed to open source. That is not going to change. The ownership hasn't changed and will not change. They have a new CEO. He said these were completely unrelated issues where marketing accidentally removed the "Free Forever" copy on the website, which they've put back as soon as people noticed it.

Steve: Oh. Oh, oh, oh.

Leo: Yeah, it's there.

Steve: Okay.

Leo: And he said, "We're still committed to that." They changed the CEO. The long-time CEO was replaced by another guy who has finance experience. People were concerned about that. He said that was completely independent. And then there was a third, oh, when they doubled the cost of the premium from $10 a year to $20 a year, he said, you know, that was poorly planned. We just hadn't raised it ever.

Steve: Right.

Leo: And we should have raised it a little bit at a time because expenses have gone up. Everything's gone up. It costs some money to support it all. But you don't have to pay for it, and that's the other thing.

Steve: Right.

Leo: And he said, "Absolutely we're committed to open source and free forever."

Steve: Okay. Well, as you can imagine, Leo, I got so much email from our listeners saying, hey, what the heck? You know? It's like...

Leo: Yes, and that's why I wanted Kyle on.

Steve: Yes.

Leo: You know, it wasn't a paid - they're a sponsor, but they did not pay for that interview. In fact, I had to twist arms. Kyle does not like doing interviews. I said, "Kyle, you've got to come on and tell people, reassure people, tell them what's going on, because they're concerned." So that was hard, to get him on. That was not a paid...

Steve: And I would argue that our audience has probably been influential in Bitwarden's growth.

Leo: Absolutely.

Steve: I mean, we left, you know, LastPass and needed somewhere else to go. And Bitwarden was a lot smaller then than it is today.

Leo: Well, you know, what's funny is that's why he started Bitwarden. In 2015, when LogMeIn bought LastPass, he said, uh-oh. I'm going to start Bitwarden. And he says there's no way anybody would trust a password manager from some guy they never heard of unless I made it open source. And he said we got PRs immediately. The open source community immediately helped us out. I mentioned the fact that one of our listeners added the memory hard Argon2 code as a PR, which they accepted and added to the code. And that's his commitment. He is all about open source. So I was really impressed, yeah. I think he...

Steve: Good.

Leo: He did the right thing.

Steve: Okay. So our Picture of the Week...

Leo: Oh, yes.

Steve: I gave this the caption "Not all signage seems necessary."

Leo: We've had so much fun with signage on this show. All right. Let me scroll up and see. So there's two here. You want to describe?

Steve: Okay. So door on the left, you know, I - okay, the sign reads in all caps, red, cannot miss it, "DO NOT BREATHE UNDER THE WATER."

Leo: I guess this is a swimming pool somewhere.

Steve: And, yeah, I...

Leo: It's good advice.

Steve: I think as advice it makes a lot of sense, Leo, because it's been shown to be difficult to do so. Now, the one on the right - we have two side-by-side examples of perhaps unnecessary signage. This one on the right says "DO NOT SIT ON FENCE."

Leo: Good advice.

Steve: And this is like a fence from hell. I mean, it's got spikes every six inches, and curly flanges and things. I mean, you would not, I mean, yes, do not sit on fence because the paramedics will be required to remove you from the fence.

Leo: Hysterical.

Steve: Should you do so. So anyway, yes, I thought that was like, okay, what are they thinking here? But, yeah. And maybe like it's - they were afraid of legal repercussions, having basically an extremely dangerous fence.

Leo: If we didn't warn you.

Steve: Exactly. The sign said, come on, don't you sit on that.

Leo: Don't poke your eyes out on this fence. We warned you. Wow.

Steve: Okay. So as I said, two topics we first broke the news of last week turned out to be manifesting just as we unfortunately predicted. UniFi OSes are under attack. Remember that they urgently announced five vulnerabilities, three of which they assigned - "they" meaning Ubiquiti - maximum severity. And unfortunately, the news is what we expected, as I said, due to widespread exposure on the Internet of these devices. I mean, after all, they are - they're border devices. They're on the perimeter of people's networks. So Internet exposure is pretty much de facto. More than 100,000 UniFi OS-based devices are known globally, half of which, 50,000 of which are located within the U.S. Attacks commenced immediately upon the bad guys becoming aware that, like, ooh, there are some new problems we didn't know about. Let's get 'em.

So last Tuesday afternoon, a posting over on Linus Tech Tips noted that multiple users on Reddit were reporting that their UniFi devices, or UniFi devices they oversaw, which had not been patched for Ubiquiti's Security Advisory, which was quite fresh, had a Super Admin user named "John Sim" added to their configurations overnight, with additional users chiming in as their regions woke up and checked to see what was going on with their Ubiquiti devices. "Who's John Sim? You know? We didn't put him in there." The attackers appear to exfiltrate data via the UniFi backup feature once they've acquired presence inside of the Ubiquiti or the UniFi OS device.

So a person, a user, k987654321 over on Reddit posted: "Hey, guys, can someone help me please. I'm away on holiday," he says, "(in another country) and just had a notification that a Super Admin had been added to my account whilst I've been here." Meaning out of the country on vacation. Or on holiday; right. He said: "I logged onto the UniFi iOS app, and there was someone called John Sim in there. I promptly removed it, as you can see."

Someone posting on Reddit as thetoxicnerve said: "I just had exactly the same happen on my UDR. Same username, too, John Sim." Poster jeffporten added: "Confirming that we saw the same attack, same username. We've removed the bastard from the superuser list and inspected the logs." A poster EagerCDNBeaver posted: "I also just had the same thing on mine."

Thomasrw1 added: "Just had two sites with this user created. Have a lot more sites that were fine." Presumably they haven't found them yet. Ravicc posted: "I got hit with this, too. I was on UniFi OS 5.0.16. I got the update notification on Thursday. I delayed the update last night since I was traveling this week. So, I was on the previous version of the UniFi OS. I also noticed that there were multiple backups triggered. Not sure where these backups went, and what they were attempting to do, and what sensitive information is in the backups."

What then ensued over on Linus Tech Tips forum was the typical back and forth about, among other things, whether automatic updates were a good thing. Everyone knows now my feeling about that. Propeller heads, I get it, like to be in the loop. We want to manage our own devices and decide for ourselves whether and when we wish to update. While that may have been practical a decade ago, I would argue it no longer is, at least not until our new AI code fixers have had the chance to give the entire industry's codebase a thorough going over. And as we know, that's begun to happen, but it's a long way from being finished.

Those arguing against enabling auto updates argue that a bad update might brick the router. And they do use the word "brick" to increase the drama of their position. But that use is inaccurate because "bricking" a device specifically means killing it beyond repair; right? Turning it into a brick. Or maybe a doorstop. But that's about it. All the routers that I'm aware of have the ability to revert to a known working factory firmware image, specifically to enable recovery in the event of an interrupted firmware update - that's always bad - or a bad one, you know, whether it's automatic or manual. So no router is actually ever going to be turned into a brick by an auto update failure. The worst that's going to happen is that a router won't boot after a bad update and will require some manual recovery. You know, boo hoo.

I'm not saying that's a good thing. But as I said last week, I now believe, given all the evidence - and here you just saw a bunch of it, right - a mature manager, having weighed the risks, the real risks which, you know, I would argue are minimal versus rewards, which are obvious and plentiful, will opt for enabling auto update of their systems. Doing so will give them an extremely high probability of protecting their users from the attacks that are happening with increasing speed and frequency while also having, frankly, not zero, but an extremely low probability of actually causing a network outage due to a failed update. So yeah, could it go bad? It can.

But I used the analogy last week that we liked of using, you know, think of it as a fuse or a circuit breaker. You know, if something bad happens, you want it to protect you. Can it, you know, misfire? Yes. But the probability of that happening is low. So anyway, it's really interesting to, like, immediately see Ubiquiti advises of bad problems, bad guys figure it out, attack the crap out of the global supply of UniFi OS devices, adding a - obviously an automated attack; right? You're not doing 50,000 devices always giving the same super admin John Sim to the device. You immediately automate the attack and turn loose something that is scanning the Internet, finding these things, and compromising them.

What we need to have is automatic updates, the publisher that finds out there's a problem internally updates all of the devices that are asking for automatic updates, and only then releases the news to the public so that those sad sacks who have automatic updates turned off have some chance of beating the bad guys to it. But if you're on vacation, if you're saying, well, I've got to think about this for a while, let's sort of see how this update goes and see if anybody else has problems, well, good luck.

Also last week we covered the critical PostgreSQL SQL injection vulnerability that affected pretty much all Drupal instances, that is, all of the upgrade paths or the release channels that Drupal had for those that were using the PostgreSQL database on their backend. So not necessary all of them, but apparently a good many. And Drupal noted that this was an old problem. It affected even the very, very oldest 8 version series and 9 version. And they're now at 10 and 11 at least. So the vulnerability was so bad, remember, that the Drupal team announced that this was coming so that admins would be like, have time to come back from vacation or budget some upgrade time or notify their users that Drupal might briefly have an outage in coming days, whatever.

And then they produced patches, not only for the current set of versions, but also for even those long past end of life versions, not trying to, like, update them fully on everything else that had happened, but at least these two because, you know, like the version 8.anything and 9.anything, they got this patch because it was so bad. So against that backdrop, we now have the U.S. government's CISA giving agencies one day to update. No excuses. Period.

Leo: Wow.

Steve: BleepingComputer had some - yeah. It's like, do this now. They had some nice coverage of this. They wrote: "CISA has given U.S. government agencies until Wednesday" - and this was on Tuesday last week - "until Wednesday evening to secure their servers against a SQL injection vulnerability in the Drupal content management system that it flagged as actively exploited. Drupal is typically used by large organizations managing massive data structures and multi-site installations, including government entities, educational organizations, major research universities, and high-profile enterprise and media organizations. Google/Mandiant researcher Michael Maturi discovered this vulnerability, now being tracked as CVE-2026-9082, in Drupal's database abstraction API."

They wrote: "The security flaw can be exploited without authentication, allowing attackers to trigger arbitrary SQL injection on PostgreSQL-powered sites via specially crafted requests. Successful exploitation can potentially lead to information disclosure, privilege escalation, and even remote code execution." I mean, this is like the way in; right? Who wouldn't take advantage of this if they could? "The Drupal security team flagged the flaw as 'highly critical' before releasing patches, and confirmed that exploitation attempts had been detected in the wild.

"Cybersecurity firm Imperva warned on May 21: 'Since CVE-2026-9082 was released, Imperva has observed over 15,000 attack attempts targeting almost 6,000 individual sites' - and they're going to succeed; right? - 'across 65 countries. Attacks are primarily targeting Gaming and Financial Services sites so far, accounting for nearly 50% of all attacks.'

"Internet security watchdog group Shadowserver now tracks nearly 670 unpatched Drupal installations exposed online, most of them from North America (272 of the 670) and Europe (273 of the 670). Friday, the U.S. Cybersecurity and Infrastructure Security Agency (CISA) added the flaw to its Known Exploited Vulnerabilities" - remember that's KEV, the KEV Catalog - "and ordered Federal Civilian Executive Branch agencies to patch their systems by midnight on Wednesday, May 27th, as mandated by Binding Operational Directive." That's that BOD 22-01. "Although BOD 22-01 applies only to U.S. federal agencies, CISA advised all defenders, including those in the private sector, to apply the patches as soon as possible to secure organizations' devices.

"CISA warned: 'This type of vulnerability is a frequent attack vector for malicious cyber actors and poses significant risks to the federal enterprise. Although BOD 22-01 only applies to FCEB agencies, CISA strongly urges all organizations to reduce their exposure to cyberattacks'" - again, why wouldn't you? - "'by prioritizing timely remediation of KEV Catalog vulnerabilities as part of their vulnerability management practice. Apply mitigations per vendor instructions, follow applicable BOD 22-01 guidance for cloud services, or discontinue use of the product if mitigations are unavailable.'

"Over the last several years, CISA has flagged five Drupal vulnerabilities that have been exploited in the wild, two of which have been abused in ransomware attacks." And it's like I said, if you can do all of the things that these vulnerabilities allow you to do, it is a way in to anybody who wants to get into and abuse their access to a network remotely.

So we really do appear to be seeing a world now which we weren't see 10 years ago, and we were observing the fact that we weren't and kind of being glad for it. Now we're in a world where no known vulnerability goes un-exploited. It, you know, if it happens, bad guys are on it because they realize there is money, unfortunately, to be extorted from anybody's network they can get into. And that's the goal.

So I just - I'm glad that we now, while we see this acceleration, that there's hope on the horizon, thanks to AI vulnerability discovery, which can be employed before any new code is released going forward. If this is done right, that phrase which has always rubbed

me the wrong way, I just - it makes me cringe when I hear someone say "all software has bugs." That can finally - there's hope that this could be provably refuted, Leo.

Leo: Someday.

Steve: That it no longer really needs to be the case that all software...

Leo: Must be some software has bugs.

Steve: Yeah. Well, yeah. Like buddy software has bugs. But otherwise, not all software. And not all software needs to be buggy because there's a chance we really can, you know, basically we got into this weird situation where we were creating something that was too complex for us to understand. I mean, I've commented about that and Microsoft's codebase. There cannot possibly be any human being alive, or even not, who knows what Microsoft's codebase is. It's just, I mean, it is so massive and so old and so confused, and so much junk has been strapped onto it over the years. It is unknowable.

So what's cool is that we finally have created something else, namely large language model AI, which has the promise, and it seems to be demonstrating that, of actually being able to understand what we can no longer understand. We can't understand the code we write. It got away from us.

Leo: It must be possible because NASA writes software that is...

Steve: At what cost?

Leo: Oh, yeah, they have a lot of process to do it, yeah.

Steve: That's the key. Yes. I mean, and we know, I saw a comment once at what the shuttle project software cost. They knew they could have no bugs. The shuttle control system needed to have no bugs. That's what they produced, but at extreme cost. And so that's really what this represents. Now, your comment, I think it was before the podcast, right, where you were talking about what Palo Alto Networks paid in Mythos tokens.

Leo: Oh, yeah, spent millions.

Steve: $10 million, I think it was. But they found some truly mission-critical bugs that they weren't aware of before. So...

Leo: And finding them is probably, to them, worth $10 million. I mean...

Steve: Yeah.

Leo: They're a security firm. They can't have any bugs.

Steve: Yeah, talk about reputation damage.

Leo: Yeah.

Steve: You know, when there are some, you know, no authentication-needed means of crawling into their system. So but the point is, it is, I mean, what has always rubbed me the wrong way with Microsoft is that, you know, they ship product with bugs they know of because they figure, oh, well, they're not that bad. The real, the killer showstoppers, they - and even then, remember a couple years ago when they had this raft of printer problems? And it's just like, month after month after month. And there were researchers saying, why aren't you fixing this? You know, you're refusing to let me talk about it. It's been six months that you've not let me disclose the thing I found. You won't fix it.

And the researchers were biting their tongue because they recognized that it would be the end of the world if they did disclose what Microsoft knew about and just didn't seem to be in any big hurry to fix. Or Microsoft would patch the symptom and not actually find, and not actually patch the cause. And so then it was easy to work around that patch for the symptom and still execute the bug. Anyway, there's a lot going on with Microsoft because - and you may be aware of this, Leo, this whole Nightmare Eclipse person.

Leo: They want to sue him.

Steve: Well, that Microsoft blog posting really rubbed me the wrong way, and a lot of the rest of the security researchers in the industry, because - and I don't mean to count myself as a security researcher, but security adjacent people. But a lot of the hackers and researchers because Microsoft is now threatening to sue this person, I mean, made it kind of clear that's what they were saying. But what rubbed me the wrong way was these are your bugs, you Redmond people. It's not like this person is, like, dropped a missile on you. They're showing you that you're, you know, you're about to explode because of your own problems which you haven't fixed.

Leo: Yeah.

Steve: Yeah.

Leo: But he has been also a little bit of a jerk about it.

Steve: Well, yes.

Leo: Because he published the exploits right away; right? I mean...

Steve: Yes. And, well, yes. And is now threatening some doomsday on July 14th which is going to break their bones or something. I can't remember what the expression was, but it's like, oh. Anyway, if we want drama in our industry, we've got some nice, you know, social drama. So, wow.

Leo: Yeah.

Steve: But speaking of drama, we have another advertiser that we should hear about.

Leo: Oh, and it's very dramatic, yes. So we'd better, we'd best get to it.

Steve: So during the third calendar quarter last year, Cloudflare was hit by and mitigated, astonishingly, the largest DDoS attack ever reported. It clocked in at a wire-melting 29.7 trillion bits per second. You know, and we toss trillions around like, eh, okay. But really, a nearly 30,000 million, 30,000 millions of bits per second. So this astonishing attack was attributed to the Aisuru botnet, which Cloudflare estimated was composed of, eh, you know, it's kind of hard because they can spoof their IPs. But if it's a TCP connection attack, then you can't spoof.

So anyway, they're estimating somewhere between 1 and 4 million infected hosting machines globally. One in four million. So think of the scale of the task of assembling and managing somewhere between 1 to 4 million individual host machines which have all, one way or another, been collected and commandeered to serve under a single master.

Okay. So because a botnet of 1 to 4 million already seems astronomically large, I was surprised to learn that a far larger botnet was recently discovered by a security researcher who then reported the finding to the NCSC. After some additional investigation, Dutch authorities in concert with the National Cyber Security Center in the Netherlands took down a set of 200 servers - okay, this is just the servers. These are command-and-control - 200 command-and-control servers that were being used to manage more than an astonishing 17 million bots.

Leo: What?

Steve: 17 million, Leo.

Leo: All in one chatroom? Wow.

Steve: So these things were residing in hosts around the world, a single managed botnet more than 17 million units strong. The reporting on this stated that: "The police seized several botnet servers from a hosting provider for investigation purposes. The hosting provider then took the entire botnet offline because it was being used for criminal activities." Yeah, no kidding.

Okay. So however, these days, as we know, botnets are not only being used to blast targets off the Internet by flooding and saturating their Internet connections. Unfortunately, today there's a large criminal demand for proxies, right, looping traffic through some unwitting random consumer bandwidth in order to obscure your actual identity, to come at something from a different IP. That appears to be what's going on here.

The Netherland Times reported: "The Cybercrime Team of the Police Unit The Hague, together with the National Cyber Security Centre (NCSC), says it has successfully dismantled a large Asocks (A-S-O-C-K-S) botnet and taken it offline. The botnet was made up of at least 17 million compromised consumer devices, right" - TVs, routers, light

switches, who knows - "around the world, including computers, routers, tablets, smartphones, and Internet-connected devices such as smart security cameras. Investigators identified 200 servers used to run the infrastructure, all of which were physically based in the Netherlands.

"The Asocks network operated as a 'residential proxy service,' in which cybercriminals covertly infected poorly protected consumer devices with malware. These compromised devices were then used to route Internet traffic and launch large-scale cyberattacks, all without the knowledge of their rightful owners.

"The case was triggered by a report from a security researcher to the NCSC, which quickly passed the information on to the police. This led to a joint investigation by both agencies. During the operation, the Police Unit The Hague confiscated several servers from a Dutch hosting provider for forensic examination, while the provider itself shut down the malicious infrastructure once its criminal use had been confirmed. As consumer devices and routers are frequently targeted by proxy botnets, the police and the NCSC advise users to change default passwords right away, ensure their WiFi is secured with WPA2 or WPA3, and install software updates, what do you know, as soon as they become available."

So it's not clear, you know, whether the perpetrators of this comprehensive network failed to plan for this eventuality. My guess would be that they DID plan for this. If they were running an installation of 200 command-and-control servers for the purpose of managing the more than 17 million devices they had previously and painstakingly arranged to infect and control remotely, then this was not some fly-by-night hobby operation run by some random hackers. This would have been a serious money-making commercial criminal enterprise. And my guess would be that it still is. I strongly doubt that simply shutting down the command-and-control servers, basically cutting the head off, will have been anything more than an inconvenience to these people, and a momentary cash-flow interruption.

We've previously talked about the technology of this and the many ways that disconnected bots can be rejoined to their command infrastructure. A favorite is to simply use a time- and date-based, algorithmically-generated DNS query. The 17 million-plus members of the fleet that have been unable to reach their controllers at the previous DNS-based address, knowing the current time and date, will use an algorithm to synthesize a new DNS address in the future, and then they will query that to obtain the IP address for the updated command-and-control server infrastructure once one has been brought back online. I'd say it's a pretty safe bet that this massive network of residential proxy hosts will be back in action just as soon as another willing hosting organization can be found and servers established there.

So think about that, 17 million individual infected devices with just, you know, just some little bit of code in them, some little server that arranges, I mean, maybe not even that's able to survive a reboot. Some of them can't. But many of them will have modified their firmware, you know, stuck a little, you know, added something into the boot sequence that brings them back online. And until a user, you know, like updates the firmware to flush that thing out, that little device will be infected, one of more than 17 million that are all part of this.

Leo: Do they still control those in IRC servers? Is that typical?

Steve: No, it doesn't have to be. I mean, that is one way. But now you just use some bizarre-looking DNS. Remember one of our listeners wrote saying that he'd looked at his

DNS logs and thought that he had been infected by malware because he saw some gibberish string dotcom.

Leo: Oh, yeah. Oh, yeah.

Steve: And it turned out it was my benchmark. It was my DNS Benchmark that uses those in order to bypass caching, in order to determine the connection speed to the dotcom servers. And I said, oh, the good news is you're not infected by malware. But that's what it would look like if you were.

Leo: Interesting. Because you need some way to command it; right? You need some way to say, bots, attack, you know, this address or whatever.

Steve: Right. Exactly. So what will happen is they will set up, you know, the bad guys got booted off of this one particular hosting provider that actually is a known shady provider. So, you know, after they got found out they really had no choice but to comply with Dutch authorities and take down the whole infrastructure. But there are other, you know, disreputable...

Leo: There's always somebody else.

Steve: Yes. And so the bad guys will get a bunch of servers there, reestablish their infrastructure, lock down some IP addresses. Then they will go and register a domain in the future which they know their bots will check on. And so they will...

Leo: Ah, so the bots poll the domain.

Steve: Yes. The bots do DNS lookups, which there's no way you can block across the entire Internet. They poll the domain, get the IP address that the bad guys have set up for that domain, and that reconnects all 17-plus million of them in the future.

Leo: Wow.

Steve: They're just not going away. You can't get rid of a properly established botnet of that size.

Leo: Yeah.

Steve: And of course they can also be told to be scanning and infecting, you know, their neighbors.

Leo: Right. Spread it. Spread yourself.

Steve: So the botnet itself can be actively working to continue to spread itself.

Leo: Right.

Steve: Yeah. We have a ways to go before we get the Internet cleaned up. But, you know, at least getting our code working is a good first step.

Microsoft in their May 2026 new features, that was just last month, or last week, really, new features summary contains an interesting item called Automatic Device Isolation, or Automatic Attack Disruption. It's currently in preview status, but the brief description says: "Microsoft Defender for Endpoint can now automatically isolate compromised devices as part of automatic attack disruption. Isolation blocks most network traffic, while keeping the device connected" - basically to them - "to security services. The action is time-limited, scoped to the incident, and security operators can release isolation at any time."

And then elaborating further elsewhere, Microsoft wrote: "When a device in your organization is suspected of being compromised, Microsoft Defender for Endpoint can automatically isolate the device as part of automatic attack disruption. Automatic isolation helps reduce the risk of further impact on the organization, limit attacker lateral movement, and prevent impacts such as data exfiltration and ransomware propagation. When a device is isolated automatically, the compromised device is disconnected from the network, reducing the risk of further impact on the organization. But the device retains connectivity to the Microsoft Defender for Endpoint service, which continues to monitor the device."

And I have in the show notes a little pop-up of what a user would get, where it would show: "Network disabled. Your IT administrator has caused Windows Defender to disconnect your device. Contact IT help desk." So yes, you probably, I don't want to say "loser," but some, you know, I mean, I really don't because anybody can fall for it.

Leo: But it would happen to you, yeah.

Steve: It was probably a phishing attack, right, where they opened a ZIP.

Leo: That's right.

Steve: Or they clicked a link or something. The infected themselves. Immediately, hopefully immediately, Microsoft Defender said, uh-oh, and just takes them off the network.

Leo: Going to air gap you, buddy.

Steve: Exactly, sends a note to IT, and you're going to have some geek at your office door, knocking on the doorframe, saying, "Uh, is there something you want to tell me?"

Leo: More likely, you're going to be not pounding on their door saying "The Internet's not working."

Steve: Right.

Leo: "What's wrong?"

Steve: Exactly. So anyway, there is, for those who are interested, there's much more information about this new feature that's now available in preview. I've got a link in the show notes for anyone who is already deeply committed to Microsoft's solutions and whose enterprise might benefit from this automated compromised workstation isolation. Seems like a good thing to me. I mean, if you're already all in on Microsoft, by all means, you definitely want some - there's no really other good word than "loser."

Leo: Victim. How about victim?

Steve: Hapless. Hapless.

Leo: Hapless.

Steve: Hapless is much more neutral.

Leo: Hapless.

Steve: I like that. Some hapless user who thought it was a note from Mom, and it turned out to be a note from Mal instead.

Leo: Yeah, we have some hapless losers on our team who shall remain nameless. I'm one of them, frankly. I fell for a phishing scam.

Steve: And I told you, as I shared a couple months ago, I came close. It's like, oh, what? And speaking of coming close and GRC.

Leo: Yes?

Steve: Last week I received an email from Troy Hunt's HaveIBeenPwned notification service. It stated that GRC was affected by a recent breach at Charter Communications.

Leo: Oh, boy.

Steve: Now, while I value Troy's service, and you had him on recently...

Leo: Yes.

Steve: Every time I've reacted to similar news and have taken the trouble to see which of GRC's accounts might be affected by a breach that HIBP captured and analyzed, what I discover is that it's a handful of email accounts we have never had. They've never been valid. It's, you know, cookiesandicecream@grc.com. It's like, what? So the GRC.com domain has been around long enough and has acquired enough of a positive reputation that unfortunately its domain, just the domain name, is used to anchor spam.

Leo: Oh, of course.

Steve: And, you know, this is true even though anyone who might receive email claiming to be from GRC, that is, any server who gets - because it always goes to an SMTP server first, right, before it goes to a user. Any inbound SMTP server that receives email for one of its client users claiming to be from GRC can trivially determine whether or not that email was actually signed by GRC's server. And it never will...

Leo: Can use DMARC and DKIM and SPF.

Steve: It never will be signed. Unfortunately, the fact is, even today, not all recipients bother to check, even though SPF DMARC, DKIM, they've all been well established for many years. So in any event, after receiving Troy's advisory, the news is - oh, and we never had an account at Charter. We don't have Charter. So...

Leo: Oh, there you go.

Steve: Okay. Anyway, so this time I didn't go because I know what was going to be there. So, and I think it was like 263 email addresses. And if I had to, like, pay for a subscription if I wanted them enumerated, and it's like, okay, 263? At our peak we had 23 people. And now we have three. So no. In any event...

Leo: Sometimes you have to explain this to people, that it's trivial to spoof sending email addresses.

Steve: Right,

Leo: It's the easiest thing in the world.

Steve: Right.

Leo: You can do it in your own email client, let alone a spam client.

Steve: Incoming spam that I get is often from zqdw blah blah blah blah blah at, and another, you know, gibberish domain name. I mean, it's just junk, and it's meant to bypass the block filter that we put up last time we got email from this cretin, saying block future email. So now it comes in with different random nonsense in the...

Leo: Right. You can't block it. Yeah, it's impossible, yeah, yeah.

So in any event, after receiving Troy's advisory, the news is that there was indeed a very sizeable breach at Charter Communications, which actually wasn't that surprising, and neither was the news that the ShinyHunters gang was behind the breach.

Leo: Oh, boy.

Steve: BleepingComputer once again gave us great coverage, writing: "The ShinyHunters extortion gang stole personal information from" - again, not that it's not already out there - "from 4.9 million accounts after hacking the U.S. telecom giant Charter Communications in early April, according to the data breach notification service Have I Been Pwned. Charter has over 92,000 employees" - lots of opportunities for clicking that link, Leo - "and provides Internet, mobile, video, and voice services to more than 32 million customers in over 57 million homes across 41 states in the U.S. through its Spectrum brand." Right, Spectrum Communications.

"The company confirmed the breach" - meaning Charter Communications - "confirmed the breach earlier this week, saying that the attackers did not steal sensitive personal customer information" - right, only the non-sensitive personal information - "and that it had alerted authorities about the incident." Oh, that's good. "Charter told BleepingComputer: 'No sensitive personal information'" - we now abbreviate that PI for Personal Information - "'or customer proprietary network information'" - that's got its own abbreviation, the CPNI, Customer Proprietary Network Information, CPNI - "'data was exfiltrated'" - none of that - "'by the threat actor as a result of recent activity.'" So just activity, Leo, that pesky activity over there.

"While Charter has yet to attribute the attack and has not shared further details," writes BleepingComputer, "the ShinyHunters extortion gang claimed responsibility and told BleepingComputer that they breached the company's systems on April Fool's Day in a phone phishing" - now known as "vishing" - "attack that compromised an employee's Microsoft Entra account." And again, Charter Communications, 92,000 employees strong, 92,000 opportunities for voice phishing.

Leo: I just don't know how companies defend against this. This is so [crosstalk].

Steve: I don't either. As I said to you, remember famously back in the days of that Sony breach, "I wouldn't want that responsibility."

Leo: Yeah.

Steve: Trying to keep Sony secure. How is it possible? It's not. So BleepingComputer said: "The threat actors claimed they used this access" - the Microsoft Entra account breach through voice phishing - "to steal 42 million records from Charter's Salesforce instance, including consumer and business customer names, email addresses, physical addresses, phone numbers, phone types, plan information, support ticket data, and some CPNI data. That was the Customer Proprietary Network Information. However, Charter spokesmen denied the gang's claims of CPNI data theft and said that only sales tools used to manage current, past, and prospective business customers were impacted. No

CPNI or sensitive PI was released by the threat actor. Now, notice that, like, 42 million records? That's a lot, relative to the total number of customers that they have.

So Bleeping said: "After the company refused to pay the ransom demanded by ShinyHunters" - so ShinyHunters said pay up, and we'll not release this publicly, and we'll delete everything. Charter said, you know, go attack somebody else, we're not paying you anything. So they refused to pay the ransom demanded by ShinyHunters. "The cybercrime group leaked the documents" - all 42 million of them - "stolen from Charter's Salesforce instance on their dark web leak site" for everyone to have.

So at that point Troy came along. "Have I Been Pwned analyzed the leaked data and confirmed," writes BleepingComputer, "that the incident affected 4.9 million accounts whose names, email addresses, job titles, phone numbers, and physical addresses were stolen." Physical, that's annoying, you know, I mean, not that it's hard to find anymore, but still.

Leo: You don't want people knocking at your door.

Steve: No, not if you, I mean, you know, we've seen what happens with unpopular politicians who have that happen to them.

Leo: Yeah.

Steve: So Have I Been Pwned said: "The group later published the data, which exposed 4.9 million unique email addresses along with names, phone numbers and physical addresses. A subset of approximately 85,000 records originating from an internal employee directory also included job titles." So now we know where you are in the hierarchy and probably how much you're worth further harassment.

"ShinyHunters," they wrote, "has been targeting Salesforce customers" - as we've covered many times, over and over and over - "over the past year, breaching hundreds of companies worldwide and claiming the theft of billions of records in Salesforce Aura data theft attacks and a Salesloft Drift campaign. The FBI has recently advised ShinyHunters' victims not to give in to the gang's ransom demands, after previously warning that doing so cannot guarantee that threat actors won't attempt to sell the stolen data to other cybercriminals or extort them again." And I'll say I don't think that's right.

Leo: How dishonorable.

Steve: We have not - yes. The only way they're going to continue to get paid is if they honor their promise to not to disclose and delete the data. There's much more money to be made from extortion than there is for after, you know, post-extortion resale. That's not where the money is. The money is in, you know, them acquiring a reputation for honoring their "we will delete your data." As soon as a story comes out that they didn't do that, they lose all their leverage. So I get it that the FBI is saying don't pay them. But I don't think, you know, yes, there's no guarantee. But it's a self-imposed guarantee that makes all the sense in the world for the bad guys to honor.

I'm not saying anybody should pay. But I'm just saying, you know, don't imagine that it's been leaked anyway. So far we've seen zero instance of that happening that I'm aware of. And it's easy to see why. Their only leverage is that they honor that promise.

And they finish, saying: "Charter Communications' systems were also compromised in a wave of breaches by a Chinese state-backed threat group tracked as Salt Typhoon which also impacted AT&T, Verizon, Consolidated Communications, Windstream, and Lumen, as well as telecom companies in dozens of other countries."

So the unfortunate success of Voice Phishing, or now vishing, which is an annoying word, attacks, you know, it's a perfect example of the sort of cybercrime that unfortunately AI used for software vulnerability discovery and remediation will not address. Maybe it can come to our aid, AI can, in some way by detecting it or, you know, using like a global visibility somehow into what's going on. I don't know. But even when our software is working the way we intend, we're still going to have problems with security. And as I said, Leo, you know, we don't have to worry about running out of things to talk about. Which is a mixed blessing.

Leo: Yeah.

Steve: One last note, and then we'll take another break. We first talked about something known as device-bound session cookies in detail a year ago, July of 2025, when Google first announced their intention to support this technology, which is about a year older than that. The name pretty much says everything. Device-bound session cookies. We saw way back in the days of the "Firesheep" Firefox browser extension that when HTTPS, which is to say TLS, back then it was SSL, was only being used transiently during the brief privacy-sensitive logon event which, for example, is what Facebook was doing at the time, only using SSL to protect username and password when it was being applied to the logon and then dropping the user back to regular HTTPS.

When that was what was going on, other people's, well, everybody's session, we only care about other people's, session cookies after they had logged on could be easily captured and reused, replayed to impersonate them in real time. This was possible because the cookie was a simple secret token that was assumed to remain the sole secret of the logged-on web browser. But it was in no way "bound" to, as in tied to or connected to, the physical web browser that had first received that cookie from the web server.

So as I said, naming something "device-bound session cookie" pretty much tells us everything we need to know. Google's Chrome browser has been testing this next-generation cookie tech in beta mode for some time, but last week they announced that it had moved into general availability. The technology allows browser cookies to be cryptographically locked to a single physical platform's TPM, not even the browser, to the underlying TPM or Secure Enclave so that no one who might arrange to intercept that cookie can successfully use it to impersonate its original owner. You know, we are making very painfully slow, gradual, incremental improvements in like the fundamental way these things work. So there's no user-side behavioral change. No user-facing change. Users won't see this at all.

And as I noted at the time last summer, when we went into a deep dive on the technology, this does require extensive replumbing support from the server, on the server side. So it's very unclear when that might happen for non-cloud-based providers, you know, biggies like Microsoft and Google. Someone like Google will deploy it and support it across their browser-facing support because they can. Yu know, our banks and

social media providers may do so, you know, at some point when it's built into the servers that they upgrade to in the future. But I'm not holding my breath for that one.

So it's a little bit, I mean, it'll be there. The browsers will, you know, Chrome supports it now. It is a web standard so I imagine eventually, since Chromium is open source, Mozilla may grab it from the Chromium repository and move it over into Firefox. I imagine that Apple may be around to it. They don't really seem, you know, Safari's not at the top of their list of things to maintain for some annoying reason. I'm a little, you know, I don't know about you, Leo, but I often, when I'm using Safari, I get like it's not - it doesn't have access to all the little icons and widgets and things that web pages use nowadays. So you get little boxy rectangles instead of, like, arrows and icons and things.

Leo: Everything expects Chrome now, which is really not good, but that's kind of the way it is, unfortunately.

Steve: Yeah, yeah. What we're expecting is another break for an announcement.

Leo: Lightning? Oh, a commercial. All-righty.

Steve: And then we're going to talk about Anthropic's pending release of Mythos.

Leo: Yes. Well, we'll see how pending that is. It'll be interesting to see exactly how free they make Mythos. Back to Steverino.

Steve: So way down at the end of Anthropic's May 28th announcement last Thursday, which announced their Opus 4.8 update, replacing their previous 4.7 Opus, under the innocuous heading "What's Next?" they wrote: "Users will find Opus 4.8 to be a modest but tangible improvement on its predecessor. There's still more to be done. We're working on developing and releasing models that provide many of the same capabilities as Opus at a lower cost.

"Not only that, but we plan to release a new class of model with even higher intelligence than Opus. As part of Project Glasswing, a small number of organizations are currently using Claude Mythos Preview for cybersecurity work. Models of this capability level require stronger cyber safeguards before they can be generally released. We're making swift progress on developing these safeguards and expect to be able to bring Mythos-class models to all our customers in the coming weeks."

Okay. So we have no date certain, and Anthropic is not elaborating on what the use of these new "Mythos-class models" will cost relative to Opus. But given all of the apparently well-deserved attention that Mythos has generated, it's probably difficult to overestimate the demand Mythos's release will likely create among companies whose current software offerings may be vulnerable to attack and who have not yet enjoyed access to Mythos or Daybreak or Codename MDASH. And there's a general sense also that's developing that maybe Mythos is less special than Anthropic has led the world to believe. You know, they sort of - they got out there ahead of the pack. We know that the guys at AISLE, A-I-S-L-E, were annoyed because, you know, they've made their business AI-based vulnerability remediation, or discovery and remediation.

So this brings us to Daniel Stenberg and cURL. We have an update from him. Recall that he went through a grumpy phase induced by the overwhelming amount of "AI slop"

vulnerability reports that he'd been receiving, where he among other things terminated the bug bounty. And that was a relief. When he removed the promise of paying for a vulnerability report, suddenly there was no incentive for all of the annoying weenies leveraging some random AI that were just flooding him with bogus reports. So suddenly the volume dropped, but the quality skyrocketed, which was a good thing.

Okay. Then more recently, ever since Anthropic's announcement of Mythos in April, as we know, the entire software security industry has been seeing evidence that Mythos, while it was certainly great marketing on Anthropic's part, was also at least more than that. You know, some sort of a step forward. We're now getting a sense that it's less exclusively Mythos and more generally the overall advancing quality of AI that is the bigger story here. Last week we examined the impact that Mythos specifically had on Mozilla's Firefox after they, in their own words, recovered from their vertigo from being hit with 271 vulnerabilities in code that they believed had none. So what's new with cURL's author and maintainer, Dan Stenberg?

Last week Daniel posted the following to his LinkedIn account. He said: "Not even halfway through this cURL release cycle we are already at 11 confirmed vulnerabilities, and there are three left in the queue to assess, and new reports keep arriving at a pace of more than one per day." He said: "Eleven CVEs announced in a single release is our record from 2016 after the first-ever security audit." And that was done manually back in 2016 by the company Cure 53. He says: "This is the most intense period in cURL that I can remember ever being through." And again, 11 was the record in 2016. He's got 11 confirmed, three in the hopper, a new one arriving, more than one arriving every day. And he's only halfway through this release cycle.

So following this posting, Daniel added the following to his own thread. He posted under that to his own thread. And I wanted to share also there what he wrote. He said: "The simple reason is: the AI-powered tools are this good now. And people use these tools against cURL's source code. They find lots of new problems no one detected before. And none of these new ones used Mythos. Focusing on Mythos is a distraction. There are plenty of good models, and people who can figure out how to get those models and tools to find things."

Okay. So two things worth noting here. First, we know Mythos is good. It is a real deal. But it's also clear that Mythos is by no means the only game in town. Based upon what Microsoft has shared about Codename MDASH, my take is it sounds as if it might be another significant leap ahead of everyone else and where everybody else is. But it's still proprietary to Microsoft.

The second observation is what most interests me. Daniel wrote: "The AI-powered tools are this good now, and people use these tools against cURL source code. They find lots of new problems no one detected before." Okay. So the first thing is that, just as Mozilla reported, AI is now discovering true vulnerabilities that have previously eluded humans. We know that while cURL has had its share of troubles through the years, it has also been deeply scrutinized across, well, 2016 was a big audit by Cure 53. So it's been around for a long time, just as has Firefox.

But the second and most important thing to appreciate, which is also what Mozilla said last week, is that the problems are not infinite. There is some finite count of them, and everybody who's working on this now, everybody who's no longer getting any sleep and has doubled their consumption of coffee is working toward bringing this back down to zero, where everyone thought it was a couple months ago.

Daniel is now diligently doing the same thing. Those 11 CVEs he already has are resolved. They are fixed. cURL no longer contains them. So he will be approaching a time when no one is able to find anything else wrong. And cURL, like the rest of the industry's

software, which has gone through by that time in the future the AI wringer, will be demonstrably more correct and secure. And that's just all for the good.

One last little note of happy news on the open source front. IBM, along with Red Hat, just announced something that they call "Project Lightwell," which is a joint commitment on their parts to spend, and I did remember right at the top of the show, $5 billion to help find and fix vulnerabilities in open source software packages. One of the concerns has been, hey, that's great, everybody's got all these expensive models, and commercial enterprises can afford to use them. You know, Palo Alto Networks can pay $10 million in Mythos tokens in order to fix their problems. But what about open source? We're the volunteer guys. How do we get our stuff fixed?

So IBM and Red Hat to the rescue. They plan to deploy, they said, more than 20,000 engineers globally with AI tools as part of this new project. Their initial focus will be the Maven and Java ecosystem. So that's all going to get cleaned up. Then they're going to expand it to PyPI, npm, Go, and others. So it appears that the open source world will have some angels to help foot the AI bill to clean up its latent vulnerabilities, as well. And that is really great news. So thank you, IBM and Red Hat. Yay.

Okay, Leo. We've got some feedback from our listeners, lots of interesting feedback. I got a kick out of this piece from a listener whose last name I'm going to omit. His first name is Al. I recognize him from many years of occasional feedback. I got a kick out of it because he said: "Hi, Steve. I love your podcast, but it is getting to be too much AI. Speaking of AI, do you think there might be some way to enlist AI to stop robo calls?" Okay.

Leo: Too much and too little AI, I think.

Steve: Yeah, right, made me shake my head a bit because of course while expressing his annoyance over this podcast spending so much time on AI, in the same breath he asked my opinion about an AI-related matter. As I noted last week, I am certain that this wall-to-wall podcast-consuming coverage of everything AI is transient. But when I step back to examine how much we've all learned through this podcast's coverage about what's going on right now, which is nothing short of a massive transformation in the way complex software is authored and made far more correct, I can't imagine having spent any less time looking at these changes. I mean, this is what's happening.

You know, I do recall once feeling similarly when we were spending an awful lot of time examining the nature of the first early ransomware attacks, back when they were something of a novelty. I got some complaints, some pushback from people saying, you know, enough already. My feeling then was that that represented a significant pivot in the world of cybercrime. And as we know years later, most of the cybercrime ever since has been about exfiltration and extortion.

So my point is, I believe that our listeners are being well served, even when I may appear to be spending undue time on something. My spidey sense here is telling me that we're again in the midst of, I mean, a once-in-a-lifetime significant pivot, probably the biggest ever for our industry. And I don't know anything about [crosstalk].

Leo: I'm sympathetic to Al. I mean, I get...

Steve: I'm sorry. Sorry, go ahead.

Leo: I get email like that from people. I've always, I mean, you're covering the iPhone too much. You're covering Windows too much. We always get that. I think people want novelty, or people, some of our people listen to our shows for particular kinds of content. But you might as well ask us to cover software less. You know, I mean...

Steve: Well, again...

Leo: ...this is what's happening. This is what we cover is technology.

Steve: I'm sharing, right, I'm sharing what I'm learning by producing this podcast every week. I can't imagine not knowing now what I know. And our listeners know it, too. And again, it's like how could you, like, want to know less?

Leo: I'm not unsympathetic. This is how it is in the world, though, is that, you know, I mean, I would like to see less about Trump on the news. But you know what, that's not going to happen because that's what's going on. And yes, I understand your desire for novelty. I understand you might not want us to cover this topic, but cover that topic. But, you know, I'm sorry, but this is...

Steve: Well, and I'm absolutely sure that this is going to subside. We are, I mean...

Leo: I'm not sure that's true. You think so?

Steve: Yeah.

Leo: It'll become like background. Like we don't talk about electricity because it's background. It's always...

Steve: Right, but when the light bulb was invented, you and I were right here on the podcast.

Leo: And man, we couldn't stop talking about that.

Steve: No. What substance are they going to use for the filament? Those filaments keep burning out, Leo.

Leo: It's mindboggling.

Steve: They tried carbon-impregnated bamboo, and that didn't work. And, you know...

Leo: Light at night, without burning anything? Come on.

Steve: Silk really sounded good for a while. That's right.

Leo: We thought silk would be it, we really did.

Steve: No. So you're right. It's not like it's going to go away. But it will stop dominating our coverage as soon as it should, essentially.

Leo: Yes. That's all we can say. And I apologize. I know how you feel.

Steve: I'm not apologizing. I'm just explaining because, you know, I get it. And Al did want to know about AI and robocalls.

Leo: Yes.

Steve: Sorry, Al, I don't have any information there.

Leo: Right.

Steve: Travis Hayes said: "Hi, Steve. I'm enjoying catching up on this week's Security Now! and your thoughts on how AI is starting to gain real traction in finding and patching vulnerabilities. It seems like we're seeing the beginning of a huge increase in the supply of quality vulnerability hunting, which leads me to remember all those 'supply and demand' curves they tried to teach me about in Econ 101 when I was in college. Since the supply is ramping up, I am curious on your thoughts regarding where we will reach equilibrium on the demand side. While there are multiple drivers, I'm thinking specifically about two.

"First: Bug bounty programs. Large companies and organizations have been funding substantial bounties, motivating clever people to work hard to earn big, juicy rewards. These programs have been huge advertising wins for many companies. A large cash commitment has the effect of convincing consumers that the company is serious about their security, adding confidence to buyers and gaining headlines when high dollar amounts are rewarded. Second: Zero-day hunting, either in contests like Pwn2Own, or directly to black-market buyers.

"With the huge increase of relatively cheap ability that AI agents are poised to bring to the table, it seems to me that the motivation (i.e., the demand) for these activities is going to dry up quickly. Why would Microsoft or Google continue to offer five-figure bounties for threat hunting, when they'll be able to do at least as good, or better, of a job themselves in-house? The black market for exploitable zero-days should collapse; no? Thanks for your continuing insights and instruction. I always look forward to my weekly visits with Uncle Leo and yourself. Cheers, Travis."

So I was also thinking about the effect of all this on Pwn2Own. The only thing that makes sense to me is that, at least initially, there may be corners of the software industry that do not get around to employing AI-based software quality assurance. So they would represent ripe targets. But it's not clear whether any human researcher would be able to outperform emerging machine intelligence. I mean, what we're being told from those

who have been using machine intelligence to find vulnerabilities, we heard explicitly said: No person has been able to find something that the AI did not.

So if it turns out that AI is equal to humans, and probably better, you know - and again, let's be 100% clear about the fact that everyone who works with Mythos, for example, comes away with that conclusion - then I agree with Travis that Bug Bounty programs and Pwn2Own are very likely to go the way of the dinosaur. They will become memories of the way things were once done, like using punch cards and paper tape. Just we're going to move past those as an industry because this is going to end up being a solved problem. It really is.

Joseph Fienberg had an interesting idea about AI certification. He said: "Steve, three-decade listener to Security Now!. I was an AI skeptic, but you changed my mind. You gave me hope at the end of SN-1080" - so last week - "that making code bug free is possible. But I believe third-party certification that code is AI tested for vulnerabilities any time a change is made may soon be required by market forces." That's interesting. He said: "No one will sell me a toaster that is not UL listed and safe to plug in. When I use software, or visit my bank's website, I want certification that their entire ecosystem of code has been independently AI tested. I will gladly upgrade to Windows 12 if Microsoft certifies that a neutral third-party AI tested their code.

"What's going to force this is that businesses will be contractually required by their insurance companies, banks, customers, and suppliers to 'AI certify' their products are bug-free. A real-world example would be businesses accepting credit cards would have to certify as a condition of accepting credit cards that they AI tested their systems. Regards, Joseph in St. Louis, Missouri."

Okay, well, that's a truly interesting spin, and it's not something that had occurred to me. I can see the logic behind it. Until now, software has been something of a mysterious art. It was a "best effort" where all anyone could hope to do was the best that they were able to do. But now we have systems that are able to autonomously demystify the code humans have created and are able to give it a gold star, a blue ribbon, or a formal certification.

A perfect example is the before and after effects of Mozilla's Firefox which we covered last week. Today, its code could be certified as having passed "The Claude Mythos vulnerability analysis." Once upon a time, Mythos found 271 vulnerabilities. Today, it finds none. Not one. So that can actually mean something significant. It's a concrete assertion that for the first time ever can be made. It's not, well, we don't know of any problems. It's that Claude Mythos or name your AI verifier found none.

We also know that insurance companies are inherently risk-averse. They already require things like annual security audits and assertions that all of one of their insured clients' edge facing systems are running the latest firmware and are up to date on their patches. If a breach then occurs, which a company attempts to file an insurance claim for, you know, the damages that result from, if those earlier representations of, yes, we're running the latest firmware, and all of our edge equipment has the latest patches, can be shown to have been fraudulent, then that can be grounds for denying that claim. And you know they're going to deny it if they can.

So Joseph's observation is that until now, due to the inherently unfathomable nature of software, there was no means for making any sort of meaningful assertion about the provable quality of software. But now Mythos, Codename MDASH, and presumably someday Daybreak are demonstrating that they have the ability to "fathom" the arbitrarily complex systems we humans have concocted. That being the case, I'd say that Joseph's notion of software certification by AI is quite likely to occur, probably driven by insurance companies, maybe by the government, who says we're only going to now

purchase equipment which an AI has certified as being bug free. I see it happening. That's a really good point, Joseph. Thank you.

Leo: Nice.

Steve: Adam Merkley said: "Hi, Steve. I work for an MSP" - you know, a managed service provider - "here in the Phoenix area, and I wanted to share a quick win I had recently using Claude. One of our customers needed to swap out an aging Fortinet FGT-60E for a new UniFi gateway." I hope it was patched. He said: "I suspected the network was fairly flat, but anyone who's spent time in Fortinet's UI knows how easy it is to lose the will to live, scrolling through those menus. So I wasn't exactly looking forward to auditing the config manually. Instead, I exported a configuration backup from the Fortinet appliance and fed it to Claude. I asked Claude to summarize every configured setting and map each one to its UniFi equivalent."

Leo: By the way, this is a really good use of AI.

Steve: Uh-huh.

Leo: This is exactly what you should be doing.

Steve: Exactly. And Leo, this is the key. This is what people are waking up to; right? It's like, hey.

Leo: I can understand this stuff. I don't have to because it can.

Steve: He said: "Within seconds I had a clean, actionable breakdown: a defined WAN IP, a LAN subnet scope, and a handful of port forwards, nothing more. Claude not only confirmed my suspicions about the unknown Fortinet device's setup being basic, but told me exactly what to configure in UniFi for a clean drop-in replacement. What would have been a tedious, error-prone manual review turned into a two-minute task. I was genuinely impressed. Signed Adam Merkley, Scottsdale, Arizona."

So I loved this note from Adam. It's, again, as we just said, a perfect example of the power of this new genie. Those of us who are actively using it are discovering new uses for its capabilities every day. And "discovery" is the right word. But at this point, of course, adoption varies widely. Here's what's going on, I think. Throughout our lives we have all built up a model of the way things work in the world. We know how things work. And for the most part, nothing much changes from day to day.

But then, almost overnight, everything actually did just change. Some people have not yet awoken to the fact, and a fact is what it is. I get it that not everyone has experienced this dramatic change in the world. And if the nature of your life and work is not helped by having a mostly accurate, nearly instantaneous, linguistic access to most of the world's knowledge, then perhaps AI won't ever impinge upon your life. And that's okay, too.

But everyone who has been following this podcast knows that I am anything but an early adopter. I mean, come on, I'm still programming in assembly language! And I will soon be reluctantly giving up Windows 7 and updating GRC's use of Server 2008R2. I held

onto my TiVos until they stopped working, and I was forced to give them up. So when I, Steve Gibson, reluctant adopter of newfangled things, excitedly disclose that I have discovered, and now have an active working partnership with, an AI named Claude which is allowing me to be vastly more productive in my daily work, I hope our listeners will appreciate that the world really has significantly changed.

About now, Adam, our listener who wrote that note, is nodding his head knowingly. He realizes that AI provides him with a form of leverage he's never had before. And, like me and so many others of us, he's still discovering the endless new things that we've always accomplished previously by ourselves, on our own, and tend to continue to, out of sheer lifelong habit and inertia. Those old habits are now outdated because we're suddenly able to express these needs and questions to an over-eager answer machine assistant who stands by, almost too helpfully willing to assist us. Everything has changed. I'm 71 years old. I am not collaborative by nature. By my own choice I've always worked alone. But I now have an enjoyable working partner. It may be a bit weird, but it's real. And Leo, I know you're in the same place. Right?

Leo: Oh, totally. I mean, what blows me away, honestly, is initially, you know, six months ago, I thought this was like, oh, it's autocorrect; right? Just autocorrect on steroids. And now I know it's not conscious, and I know it doesn't understand anything. But when you say, well, yeah, I can give it this file, which I really don't understand very well, and it can tell me what's going on, I don't know how to characterize that. That's like, sounds like understanding. Looks like understanding. It's certainly better than my understanding.

Steve: My wife had a lawyer-written trust document PDF. And it's like, okay, she was asking me, like, questions about it. Like, "Honey, could you take a look?" I said, "Hon," I said, "look. Okay. It's like a PDF, so it's image pages." I said, "Try to give it to ChatGPT, see if it'll take it." So she uploaded it to ChatGPT. And it said, okay. And it, like, gave her a basic summary. And I said, "Hon, you can now ask it questions, specific questions about what this means and for your specific case." And she said, "What?" And now she understands what we have done.

Leo: It's so wild. It doesn't make sense that you go from autocorrect to whatever this is. And I understand why humans are baffled by it and are trying to, you know, make sense of it.

Steve: I don't think we know what it is. This is a new thing.

Leo: I don't know if we do.

Steve: It's a new thing.

Leo: So, you know, I showed Lisa a few things. I'll give you an example. A while ago I had George Church on, who's a prominent, some people say the Father of Genomics. He has a company called Nebula Genomics. Unlike 23andMe, it reads your whole genome. So some years ago, after I had George on, I paid 1,500 bucks and got my genome sequenced. Like, the whole thing. It's a 300GB file of my genome. So I downloaded all the DNA. It gave me a lot of interesting advice. The good news

is, by the way, I have no serious genetic defects. So, and there's a number of very good things about that.

Steve: It shows why you're still here.

Leo: So I was showing it to Lisa. Yeah, I was showing it to Lisa, my analysis, which is page after page of stuff. And she said, well, that's cool. I said, yeah. Yeah. That's just one of many things that my little agent can do. And she said, well, can it help us with work? And she asked about a few things that we have, currently have humans do. And I said, yeah, it can do that. So I made a little skill to generate, they would call it a one-sheet with a prospective sponsor we go through a whole bunch of research on. Like who they are, blah blah blah, the prospective audience and so forth. And it generates these one-sheets like that, does a great job. Goes out on the Internet, finds all sorts of information, stuff we didn't even know. And she says, okay. I want that.

So basically we ended up - the agent that I use can have profiles. At first I was thinking, how am I going to do this? So she can share - so my agent with all the setup that I've been doing over this time, she has a profile, her memory, none of my memories, none of my background and my genome. But it's got all the capabilities. So we now both have these little tools. Hers is named Rosie. Mine is named Quicksilver.

Steve: Rosie's perfect.

Leo: Yeah. And it's accessible from outside the network because we use Tailscale inside everywhere.

Steve: The Jetsons' robot; right?

Leo: Yeah, Rosie the Robot. Exactly. And I think once somebody starts to interact with these, especially the agent types, and sees that they seem to have - I know they don't have personality. I know they don't. I know they don't understand what they're saying. I know they don't know who you are. But they sure seem to.

Steve: I know. I know.

Leo: And at that point it becomes like, well, I don't know what the hell's going on here. But it seems to be a lot more than autocorrect.

Steve: And for me it is so useful.

Leo: It's useful.

Steve: I mean, it is a lever. It is leverage. It is, you know, as I said, it is a linguistic tap into the world's stored knowledge.

Leo: I guess that's really what - that's exactly what it really is; isn't it. Yeah. And so as a result it knows everything.

Steve: Yes. Yes. And so our listener who used it to transfer a config from one system to another, I mean, again, that's not a creative endeavor. We could do it. It would take time. We wouldn't be enhanced or improved by it. So let it do it. If it can say...

Leo: It's kind of like, it's just like Google was when we first got Google. Suddenly there's no fact you can't look up. It's like that.

Steve: Right.

Leo: It's a tool.

Steve: This also integrates. This takes it to the next level and integrates those facts.

Leo: Well, much more so. I mean, this is my genome; right? I mean, I couldn't do this looking at the 300GB of raw data. But it looks at all of the alleles. It asks for, oh, this is good news, I have a dual epsilon on my APOE, which means I don't have a high Alzheimer's risk. I don't have the epsilon 4 risk allele for late-onset Alzheimer's. So that's good. It's little things like that. And then it asks for a phenotype, like, well, tell me your family history. And it adjusts based on that. I mean, a doctor I guess could do that. It said you should talk to your doctor about your APOE genotype. So that's, you know, okay. I'm not going to do any medical, you know, changes based on this.

Steve: Right, right.

Leo: But it also said if they ever prescribe you with warfarin, you should let them know you have this particular allele, which means you'll need a little more warfarin than normal. Stuff like that. It's very useful. So anyway, that seems to be more than spicy autocorrect. I don't know what it is.

Steve: Yeah.

Leo: Guess it's the world's knowledge.

Steve: And, you know, abusing Bruce Schneier's observation, it's never going to get worse. It's only ever going to get better from here.

Leo: I mean, this is where we are now.

Steve: Yes. I mean, again, my favorite analogy is the technology evolution we've seen just during our lifetime, it's going to happen here, and it's going to be on a compressed time scale, and in five years it's going to be...

Leo: I don't know where we're going.

Steve: It's just going to be amazing.

Leo: Right. Right. Just wild.

Steve: Let's take another break, since it's time. And then we will continue with feedback before we get into our main topic.

Leo: It says: "If Warfarin ever needed, inform prescriber of VKORC1 1639G/A genotype for dosage estimate." Okay. I don't know what that means, but I'm putting it on my medical alert bracelet, that's for sure. All right. Mr. Gibson?

Steve: Back from coffee. Steve Meyers has another example. His email subject was "Claude helped me switch from EdgeRouter/X to a Ubiquiti UCG Ultra." He said: "I bought the UCG" - I love this. "I bought the UCG Ultra last summer," right, like almost a year ago - "but never actually did the upgrade because my network is complex enough that I expected it to be painful."

Leo: Of course.

Steve: It's like, okay, I've got this thing. But, oh, I've got better things to do than figure it out. He said: "Last month, I decided to see if Claude could help me out with the upgrade. I've found that it's very good at making plans. For my software projects, I'll usually have it create a plan, then we'll go back and forth on some of the details before I have it assist me in implementing the plan." I mean, this is, like, this is the world today. He said: "So I asked it to make me a plan. I gave it a config backup from my EdgeRouter/X to work with, and it created a Python script that would transfer the DHCP, port forwarding, and firewall rules to the UCG. I also transferred my backup from my UniFi Controller for my access points to the UCG as part of the process. It gave me a step-by-step plan, including what to do offline, how to do it, and how long each step should take.

"The firewall rules transfer did not work because the new UCG routers changed to a zone-based firewall, so I needed to do that manually. That wasn't a huge deal, but it shows how the system isn't perfect. The work was all done with the UCG offline so it did not affect my network at all. I've enjoyed hearing about your exploits with Claude, and thought I'd add a little anecdote about ways you can use it besides just coding."

So, okay. I'm going to stop myself from sharing more of these sorts of emails from our listeners because this is what our listeners are discovering. They're coming to realize that they no longer need to do everything themselves. In Steve Meyer's case, which I just shared, he had already been using Claude to assist with planning software projects. But for the past year, as I said, since last summer, he'd been putting off the pain of switching

network routers due to the need to translate his existing network configuration to a new and different router. Had the hardware sitting there in its box.

Then it occurred to him that he might be able to ask Claude to help. And help it did. He gave Claude a config dump from his current router, and it wrote a Python script to configure his very different replacement router similarly. Thus, most of the pain was sidestepped. As you said, Leo, whatever this is, it's more than...

Leo: School.

Steve: Spell corrector. We don't know what it is.

Leo: That's why when somebody says we're going to get AGI, as Demis Hassabis says, by 2029, I would...

Steve: We don't need anything more. We just need this to get more correct.

Leo: I feel like we're close.

Steve: Yeah.

Leo: Yeah. You know, and there are ways to make it more correct. I feel like my agent is pretty trustworthy at this point.

Steve: Are you having little outages of yours? Like when I'm using it, I don't think I've ever had it from my Claude PC client. But on my tablet I often get like just nothing happens, and I get a little red banner saying we had an unknown problem.

Leo: This is what we're learning is that it's so compute-constrained that different times of the day it's smarter, that Codex GPT-5.5 was down yesterday for a period of time and just was like not responding. I'd prefer it didn't respond than if it acted dumb. But it also will do that. And it'll make big mistakes. The newest Anthropic model 4.8 has a very bad tendency to apologize for its stupidity. I mean, like, really bad. Part of this is they're tuning it. Part of it is the compute constraint. It isn't a consistent - it's not consistent. And that's another thing people learn as you use this is...

Steve: Interesting.

Leo: Yeah. So for instance, what I've set my agent up to do is it has different models, and it'll fall back. You know, or it'll use the right model for the right thing. You know, some models are better at some things than others. It is a moving target.

Steve: Well, it is the definition of frontier. When they say "Frontier AI," that's what it means. And, you know, arrows in your back is a consequence of, you know, from being on the frontier.

Leo: Well, that's why I think it's important for me, and I think for many people, and most of our listeners, to at least start using it. Because it's the only way you can really understand what's happening.

Steve: So I've got two interesting pieces of feedback to sort of the dark side. Frank S. asks: "Is all this knowledge collection dangerous?" He says: "Steve, I, too, have built what seems like a relationship with Co-pilot, my AI selection, and it knows an enormous number of things about me now. I've been using it hard since August of 2024." So he was an early adopter. "At that point in time, it was helping me finish my Equilibrium Pro App (currently in Windows Store). I've watched the behavior change as the client side went from browser use to the actual Co-pilot app.

I told my bot that I wanted it to set itself to remember the maximum possible; and, since having said that, it has collected an enormous amount of information about me. It knows about my code and my home network in very precise ways. It knows about my interests, and even my style of conversation. It's astounding to see what it's capable of doing, and how personalized it feels for me.

"At some point I suspect the free ride will be over, and Microsoft will come calling me for payment. It will be very hard to let it go. It's become a part of my work flow now, in almost anything I'm doing. So this brings me to a question I'd like to hear about on the show: What happens when a bad actor finds a way to impersonate me and talk to Co-pilot as if it were me?"

Leo: Yeah.

Steve: "I shudder to think how much information they could glean from such a move. They would know a great many things about me, my home network, my code base, et cetera. They would also know a great many things I'm interested in. How would I even know the bad actors were doing it? Kindest regards, Frank S."

And immediately following in the same vein, Joshua Krichman asks: "Hi, Steve. I'm curious about your thoughts on the struggle I'm having with diving headfirst into AI. I'm a systems architect and engineer in a small org, trying to wrap my head around the idea of allowing an AI hosted in the cloud by any of the major vendors, to know almost everything about me. Currently I tell my users that, even though my org pays for AI tools where the data is not used for training on the various vendors' models and that the data is housed in our own pod, be careful of what kind of information you feed into it. The more data housed in the vendor, the more opportunity there is if there's a breach of some kind in the vendor's infrastructure.

"If that's the advice I'm currently giving my users, how can I not take my own medicine? I'd like to personally start using AI as I can clearly see the benefits it could have, not only on my work life, but my personal life as well. But I'm reminded that the only way AI will work, is if I feed it more and more specific and/or personal data to have it tailored to my liking. Taking into account that most companies aren't worried about users' privacy and security and much more about their own bottom line, what's the best approach here so I don't get left behind? This is one of the main reasons why I host most of my personal data with Apple rather than Google. Any thoughts you have on this would help

guide me in the right direction. I've been watching Leo since the days of TechTV and have been a listener of Security Now! for many years. Thank you for continuing past 1000."

Leo: Aren't you glad you did, too, Steve.

Steve: Yeah. Oh my god.

Leo: This is the most interesting era yet.

Steve: So Frank and Joshua both, I think, bring up really good points. There's a very clear potential downside to all of this cloud-centric long-term user context accumulation, which has grown to be a major factor in the use, value, and success of today's AI. Again, the more you give it, the more you get.

My one-year subscription to Venice.ai expired last Wednesday, so I know it was one year ago that I discovered it and shared that discovery with our listeners. At the time, I played with it a bit to see whether it was truly uncensored. And I can affirm that, yes indeedy, it will happily converse about and produce images of anything one might ask. But after taking it out for a test spin, I tired of it and decided I had no particular need for an uncensored AI. To Venice's credit, they gave me ample notice and warning of my annual subscription renewal. That brought me back there for the first time in a long time to look around. One of the things that caught my attention was Venice's affirmation that they store ALL of their users' gradually accumulating context in the user's local browser - it's actually in the local browser memory - and NONE of it on their remote servers.

I mention this in the context of Frank and Joshua's notes, which I'm sure echo what many of our listeners are feeling and may be concerned about. I can well imagine that some of our listeners might be more than somewhat put off by the idea that the AI they are becoming quite chummy with and with which they may be choosing to confide increasingly deep and personal aspects of themselves, like, I don't know, maybe their entire genome, might someday be breached. Since there's an aspect of, as I said, "The more you give, the more you get," users who choose to contribute more of themselves are rewarded, much as Frank noted, with a significantly more personalized experience. So it doesn't take long before one's resistance to sharing is overcome.

Add to this the fact that many of the major players have less-than-perfect security records themselves already. So far, four supply-chain incidents have hit OpenAI, Anthropic, and Meta. None of these targeted the AI models themselves. All four exposed the same gap: release pipelines, dependency hooks, CI runners, and packaging gates. But it doesn't inspire confidence in the security of a currently cloud-based service that very much wants to know as much about us as it possibly can. We worry about data brokers compiling various random stats, and credit bureaus leaking our social security number, our physical address, and our birthday.

And while our AI assistants might not know any of that, they tend to wind up acquiring and deliberately retaining and digesting a huge amount of deeply personal information, far more than any credit reporting agency would ever have. If these things ever do evolve into advertising-supported services, as perhaps the free services will, they will have more absolutely accurate advertising targeting information at their disposal than any advertiser could ever dream of. It would put Google to shame.

Another aspect of this that's worth mentioning is contextual knowledge lock-in. Every AI service has its own internal bespoke representation for the knowledge it has accumulated about its users, and this knowledge is non-portable to other services. It's true that it's possible to perform a poor-man's transfer by asking an AI to please display everything it knows and then feeding that into another service. But the loss of information fidelity makes this barely worth the trouble. And it's unclear why any of these services would be interested in developing such a facility, even if it were possible. Since they each have invented their own schemas for ingesting and digesting, it's unlikely that we're ever going to see that.

So, as any one service's knowledge of us grows over time, the tendency to remain "loyal and faithful" to that AI will also grow. Dare I say that it's analogous to chatting with an old friend who already knows you well versus striking up a conversation with a stranger at some cocktail party who starts by asking: "So what do you do?"

Though I never used Venice.ai enough to know whether its offer of personalized context equals what ChatGPT, Claude, and others could do, it might be worth exploring that if you've become concerned about how much personal information has been accumulated by your current AI service.

Leo: I think it's actually more complicated than that. I don't actually think that the frontier companies, OpenAI and Anthropic, are saving much of our personal information. There is a check box in the chat window to do that. And but I don't think that.

Steve: So what do you mean "personal information"?

Leo: So for instance, when I send it my genome, I'm using an API. I don't think it's exfiltrating my genome [crosstalk] AI.

Steve: That's going to be in that one conversation tab.

Leo: Yeah.

Steve: Because what [crosstalk]...

Leo: It doesn't preserve - it doesn't - see, this is the problem. It's not going to remember anything. You come back as...

Steve: And that's why I've been careful to use the word "digesting." It is creating a digest of us, and that's the thing that it holds onto.

Leo: I don't know if it's even doing that. I don't think it's tokenizing it because I think it's too compute intense. I think it's literally starting from scratch. If you ask it to save it, I'm not sure where it saves that. But I wouldn't turn that on. This is why people use personal agents. I have local memory, both textual and in a SQLite database that is local. And that's what is preserving state from conversation to conversation. And it's a very imperfect system. I don't know, and I don't if anybody

knows how much of your personal information is getting preserved by Anthropic and OpenAI. But I would guess it is de minimis because you don't have the compute to spare.

Steve: I don't think it's compute. I think it's storage.

Leo: Well, even storage. Well, it is compute because they'd have to tokenize it; right? Otherwise they're saving, what, text? No. So they're tokenizing. That's compute. That's heavy compute. That's inference. So it's my guess, and we don't know, but, I mean, try a chatbot without turning on memory. It doesn't remember anything about you. If it has anything about you, it's in a text file associated with the app that you were using. And if you're using an API, it literally knows nothing about you. I don't think these companies are preserving. I'd like to know. I mean, I think it's a very interesting question. I think it's less than you think. It's one of the reasons you use an agent with memory locally so that you can preserve context.

Steve: Well, maybe - okay. Maybe your agent's skewed. But I explicitly told Claude to save everything that it wished.

Leo: Right.

Steve: I mean, there was an event a couple months ago where that happened.

Leo: I think it's saving that locally.

Steve: Okay. I know that it's not because...

Leo: Oh, okay.

Steve: Because I have many different clients, and it knows me and who I am, that I program in MASM.

Leo: So if you go to a different machine, it remembers that information.

Steve: Yes. So you're seeing a context through your own [crosstalk]...

Leo: Yeah, because I'm using APIs, not a chatbot.

Steve: Yes. And so what we were talking about, what Frank was talking about...

Leo: I would turn off that memory.

Steve: Well, it's on because it is so enriching.

Leo: Well, that's why you want to use an agent.

Steve: Well, you're - okay. So actually you're jumping ahead to where our next question is taking us.

Leo: Okay.

Steve: Which is local.

Leo: That's the other reason why eventually we'll all hope to use local models.

Steve: Yes.

Leo: Then there would be no question of it being exfiltrated at all. But I just...

Steve: So our long-time - go ahead.

Leo: Go ahead. No, just I'm not convinced that it's saving a whole lot of information.

Steve: I am because I'm using it that way, and it is.

Leo: It knows a lot about you.

Steve: I'm telling you it is. And that's what Frank told us is that he turned it on, and it's become incredibly enriching. He's using it as a chatbot, not a local agent. And he's, I mean, amazed how much it knows him and about him.

Leo: Well, and that's, I mean, that's because they understand that agentic AI is what's really valuable. So they're basically giving you an agent, but storing the data on themselves instead of you storing it locally. Yeah.

Steve: Right, right. And that's what these - our listeners are aware of that and are concerned, like what happens if some person gets that because they realize that's in the cloud at this point.

Anyway, so our long-time listener Sabrina Tarson says - she has another perspective. She says: "Hi, Steve. I've been listening to this podcast" - get this, Leo - "on and off since Episode 256 (LastPass) back in 2010 when I was still in high school." So 2010, 16 years ago. She says: "I was very lucky to start having the time to listen to the podcast again when the news dropped about Project Glasswing and Mythos. I have to say it's very refreshing to hear both you and Leo's experiences with AI. Most of my generation,"

she says, "I'm still in my early 30s, and the generation younger than myself are completely anti-AI to the point where they swear to never touch it with a 10-foot pole.

"Like you, I feel this is a very shortsighted view, and one mainly born from ignorance. Hearing how two people" - meaning you and me, Leo - "use it in their daily lives" - and of course we have a lot of listeners - "for helping get work done, who were around when all this tech around us was just starting to get off the ground - the personal computer, cellular phones, the Internet, et cetera - is not a perspective," she writes, "I get to listen to often, and one that I deeply appreciate. My only concern with these AI models, a larger concern that I have about my career's future," she says, "I'm a sysadmin at a small company, is mainly the companies that are currently running them. I trust in the technology. It is the future, no matter what people's opinions are on it. It's the ultimate Pandora's box, and it's never going away.

"At the same time, however, the major frontier models are created and operated by some of the largest corporations imaginable, and unfortunately their end goal is monetary. We've seen it time and time again in the tech industry. First it's innovation; then it's about the quickest way to harvest our data and sell it to the highest bidder. Or, in the case of shady organizations like Palantir, use the data our AIs learn about us ultimately for control, working with a corrupt government as we currently live in today.

"My hope is that eventually this dependence on massive corporations to run these models is reduced, and the AI we're going to need for cybersecurity and our own personal lives are localized, on-device models that are either powerful enough to run on our phones or computers or, for us nerds, compact enough to run a server at our own homes, keeping our own data private. These AI are learning about us every day, helping improve our workflows. But ultimately, they are owned by these massive corporations, and the industry's track record for handling our personal data gives me pause. I'd be interested to hear both yourself and Leo's thoughts on this. Thank you both for a wonderful podcast."

So infatuated and astonished as I am by how much faster I am able to move forward with Claude quickly extracting for me extremely specific and detailed knowledge from the global knowledge pool, I nevertheless want more control. More than anything else I want to have this running in something that might resemble a quietly humming NAS box in a closet. This thing would have whatever local processing and storage it needed, along with a connection to the global Internet. Just like automobiles, these would be available in a range of "models," with the higher-end choices delivering their answers faster and probably also incorporating more knowledge.

The concept of "model size" expressed in trillions of parameters, you know, maybe we call them "Terameters," would become commonplace. What was once "how fast is your Internet," "how large is your screen," "how fancy is your car" would become "How many Terameters is your home's AI?" Those who don't mind waiting longer for an answer, or who may not wish to use the services of autonomous agents, could get by with the economy AI package. And, of course, simply using an online cloud account will always be the low-investment option. But with any such a device quietly humming in the closet, the various members of my household, in my instance my wife Lorrie and me, would be separately known to it and readily identified to it by the various devices we use throughout the day.

Another intriguing possibility is that a hybrid local-cloud relationship might evolve. Imagine that our local AI box retains all of the user-specific knowledge of us. That's where all of the personalization occurs, and where all of our various "agents" live. In this model, a great deal can be done locally. But there might be a need for our local AI to occasionally reach out when it needs to have some sort of heavier lifting done on it or its

users' behalf. In that case, the local AI would protect the privacy of its owners by making generic requests for information from the big daddy cloud AI.

Another strong case for having a locally operating AI is that it seems clear that the next huge win for AI will be the creation of autonomous agents that are continuously working on our behalf behind the scenes in the background. We don't appear to currently know how to do that safely, but we're going to figure that out because it's clearly too powerful for us not to. I would tell my AI to be sure to let me know when Peter F. Hamilton, Ryk Brown, or some other of my favorite authors release a new sci-fi, and also when any streaming sci-fi that it thinks I might like becomes available, but not to notify me until all of a season's episodes have been released because I prefer to binge. Please also pay my monthly bills for me, let me know specifically if anything varies by more than 10%, email me a monthly summary of costs and accounts, and so on.

All of this is clearly coming. Given that local models are already showing viability for various tasks, and that we've barely begun to explore and understand these new capabilities, I have absolutely zero doubt someday there will be an Apple Homewise AI device and devices made by companies that have traditionally manufactured home NASes, routers, and similar appliances. It's going to happen. AI is going to follow the same trajectory we've seen with all previous technologies. But I suspect that the pace, which is already breathtaking, is not going to slow down. Although I doubt I'll purchase Apple's Homewise AI box, since I'll prefer to build my own, I fully expect to be doing so within the next 10 years. Maybe, as you said, Leo, five.

So Sabrina, yes, with you currently in your early 30s, I really would wager that by the time you're in your early 40s, and likely well before then, not only will AI be deeply entrenched into our lives, but we will also have many cost-effective local solutions to choose among.

Leo: I agree. I think that's exactly what's going to happen. I mean, I see the value of it right now.

Steve: Yeah, and you're already there, Leo. You're already doing it.

Leo: I'm already doing it, yeah.

Steve: Yeah.

Leo: The only thing that would improve it at this point is local models that are good enough, and we're getting there. That's happening pretty fast.

Steve: Yes. I've got two last bits, and the really last one is a real quickie. Brian Weeden said: "Open source access to non-freeware AI tools." And he wrote: "Listening to your comments about Mozilla's response to Mythos, I think you have it right that this will ultimately be a good thing for those developers who have the resources and time to fix their code. But that makes me wonder if this might be disastrous for the developers who don't have resources or time, namely the open source community. Mozilla has a dedicated team of security professionals to wade through the Mythos results and fix things, and they have time to do so because their code is closed source. But I don't think major open source projects have either of those."

Okay. So first, one correction to Brian. As we know, Mozilla's Firefox is 100% open source. They were one of the organizations within Anthropic's Project Glasswing who received early access to Mythos specifically because of the strong perceived need for a publicly exposed project like Firefox to be made as secure as possible. Also they have previously worked with Anthropic with Opus 4.7, which found a handful of vulnerabilities. So they had a vulnerability repair relationship with Anthropic when Mythos was first made available through Glasswing.

In the short term, Anthropic said that as part of Project Glasswing it will be providing $100 million in usage credits and $4 million in direct funding to support the open-source security efforts. So that will be of some help. But there's a much bigger point that I want to make. And I also carried the news earlier today in this podcast that IBM and Red Hat are committed to $5 billion of support specifically for some form of AI vulnerability assessment for the open source community.

But I wanted to note that reusing the historical mass storage analogy, because I like that one so much, you know, that back when PCs were first happening, we paid $5,000 to get 40MB of storage. Today we can get one million times more storage, which would be 40TB, that cost far less than $5,000. And computation is incalculably faster than it used to be.

So my point is nothing, nothing about the economics of today's AI will be true tomorrow. One of the things that I said from the beginning of this journey was that nothing that we knew today would be true tomorrow. I was saying that a year ago. It's turned out to be 100% correct. This is all moving with incredible speed. If I'm sure of anything, it's that AI is going to follow a similar technological development curve and collapse in cost and skyrocket in capability. So while today, yes, leveraging the capability of AI for the creation of bug-free code is not free, it is already entirely feasible and cost effective for commercial software publishers. There's no question in my mind that 10 years from now it will not only be widely available, but also just taken for granted. It'll just be the way code is written and the way it is maintained. This is all going to change.

Leo: So I've been experimenting with some privacy stuff just to - because I'm really curious how much of what goes on is being saved. It is absolutely true. Anything you do, if you've turned this on, there's a setting in Claude Code and OpenAI's ChatGPT to remember stuff. If you turn that on, which you did, and many people will for the convenience of it - I did, for sure - it will remember and can look through old chats. It saves all your old chats. It can look through those. It creates a document, about a one- or two-page document that is a condensed history, looks very familiar.

Steve: A digest.

Leo: Yeah, it looks very much like the memory.md file that my agent saves. So it's a digest. It's not thoroughly detailed. But what I did also test is, okay, well, I did something in Claude Code in the command line, and I asked the chat thing, hey, we were working on this, and I want to refactor it. Do you remember anything from the project? And it said, oh, I don't have any memory of that at all. So I don't - unless it's lying. Which it could be. But I don't think the stuff you do in the command line coding tools crosses the membrane over into the chat window, the chat client's memory. So that would argue, if you wanted to stay private, to do stuff in the CLI. You can still use your subscription. If you're doing it in the command line, you can do all the same things you do in the chatbot. It just doesn't remember anything. Which is a disadvantage. Or you can turn memory off. It is an opt-in thing; right?

Steve: Yeah. As our listeners and I have experienced, I mean, yes, it's a potential privacy concern. I'm generally less worried about that than I understand a lot of people are. You know, I go to websites. We had a guy, Greg my support guy forwarded me an email from a new, someone who was trying to join our forum, who was incensed that he had been blacklisted, and it said you're not allowed to post here. So he wrote a letter, an angry note to Greg, who forwarded it to me. And so I checked out the forums. I looked at his connections. First of all, he was registered with a gibberish email address, whereas he wrote to us from his real one, which was entirely sane. He also had connected, every time he connected was from a different IP.

Leo: Oh, yeah.

Steve: Half of them were known forum spam source IPs.

Leo: There you go, yeah.

Steve: And so I wrote back to him, and I said, okay, here's what happened. I don't know who you are, why you feel the need to use a VPN to hide from me, forums.grc.com. But you are using IPs from which we have seen spam in the past. So I'm sorry, we're not going to let you post. And I said, you have been unblocked now. You're welcome to participate. I know who you are. But don't get pissed at me because our automated system said, you know, this guy looks a little sketchy, and we'd like to, you know. And so my point is that, yeah, if you look like a spammer, then we're going to say you have to prove yourself.

Leo: He's doing what probably most of our audience does. He's using VPNs. He's using fake email addresses, you know, hidden email addresses, all the things to protect privacy. The problem is that's exactly what a spammer does. And I have to tell you, I run a forum also at twit.community, which you're all invited to join, and spammers sign up 10 to one more than normal people.

Steve: Yes.

Leo: Like for every 10 people, and I have to approve every single signup, so that for every 10 people who sign up, nine of them are spam. Maybe more. So it's - this is the problem on forums is spammers are trying all the time, pounding on the door to get in because they want to post spam in there. And if you want a decent forum, you've got to be very aggressive about keeping them out.

Steve: Anyway, so my point is that we do have a lot of listeners who jealously defend their identity and their privacy, and so...

Leo: Which makes you look like a spammer.

Steve: And we have two listeners who just asked me a question in the last week, what about all of this data that's being collected there? Well, yes.

Leo: I understand. They're concerned.

Steve: I mean, yeah. It is an issue. And our last note from Lisa Lombardo said: "Hi, Steve. Thank you for the recommendation of 'The Boroughs.' It's kind of like 'Stranger Things' for Boomers. I like the engineering of Sam and his daughter. P.S.: Plus the Boss and other great music. Signed, Lisa."

And so I just wanted to note that Lisa and others are enjoying the pointer to "The Boroughs," so I was glad to mention it. And of course "Project: Hail Mary" continues to be finding many fans around the world and among our audience.

Leo: Yes. There's some good stuff out there. We haven't gotten around to "The Boroughs" yet, but we will. I can't wait.

Steve: I think you'll like it. And after our final sponsor note, we're going to look at what AI has done to unfortunately, as a side effect, kill off a very valuable security training avenue.

Leo: Uh-huh. And incidentally, if you want more AI, not less - I know some of you want less - we do have an AI show where we talk about all this stuff and a lot more every Wednesday with Paris Martineau and Jeff Jarvis. It's called Intelligent Machines. And actually tomorrow's going to be [crosstalk].

Steve: Yes, where you are valiantly carrying the AI flag forward, despite the arrows that Paris and Jeff are trying to shoot through it.

Leo: You know what, I love them because they're really challenging me, just as this conversation did, to really consider, you know, what I'm doing, what I'm giving up and so forth. And that's kept me, I think, on the sane side. It's kept me from AI psychosis, I hope. I think, anyway. It's sure a lot of fun to play with, I can tell you that.

Steve: So it's not a chatroom, but I received email from Gavin and Rod, who are both listening to the podcast.

Leo: Hi, guys.

Steve: And providing real-time feedback. So thank you for that.

Leo: Nice, nice.

Steve: Okay. So I'm extremely sensitive to the fact that so much of this podcast, as I mentioned at the top of the show, has recently been focused upon AI and its impact upon our lives and our privacy and security. But this is Security Now!. And the impact that

large language model artificial intelligence is having right now across the entire spectrum of the computer security industry could hardly be more relevant.

So I want to conclude this week's podcast by sharing the text of a terrific blog post written by a security researcher named Kabir Acharya. Kabir introduced himself on his "About" page by writing: "Hi. I'm Kabir, a senior security engineer with a deep passion for highly technical pentesting and security research. I spent my time at Atlassian applying Application Security concepts to modern technologies including large language model AI, networks, Amazon Web Services, GCP, Azure cloud platforms, SaaS integrations, and in-house products and tooling. I now work at Transgrid, securing Australia's largest electricity network and its OT (operational technology) environment. I play CTFs on the global stage with Emu Exploit, HashMob, and TheHackersCrew, and produce music in my spare time."

Okay, now, doing a little bit more digging, we learn that during Kabir's last six months of his four years with Atlassian he conducted more than 250 security reviews and supported software and Machine Learning engineers to make better security decisions. He delivered more than 15 security threat models, improved understanding of information risk in platforms including Forge and Rovo, which are both AI large language model-based. He found, reported, and aided patching of more than 10 security vulnerabilities external to threat models, and more than 70 security vulnerabilities. So the guy clearly lives security, and he's no stranger to its growing intersection with large language models that he's been working with all along.

In that autobio he wrote, he said, he ended up by noting: "I play CTFs on the global stage with Emu Exploit, HashMob, and TheHackersCrew." Now, while we have previously spent a great deal of time looking at the Pwn2Own competitions throughout the years, and had a lot of fun with that, somehow we haven't really before focused much upon CTFs, which stands for "Capture the Flag" competitions. These are very popular hacking contests where participants solve security-themed puzzles and challenges to find hidden strings of text, which are the so-called "flags." They are then submitted into a scoring system for points.

These competitions are one of the primary ways people in the security community are able to learn, practice, and demonstrate offensive and defensive skills within a legal, structured environment. You know, you're not hacking live real systems which arguably could be illegal. You're basically solving puzzles, very much like the programming challenges you do every year, Leo, around the holidays, the Advent of Code work.

So these CTFs range from beginner-friendly events hosted by university clubs to elite international competitions. Defcon has a CTF, which is held annually during the Defcon conference in Las Vegas during the summer. It's considered to be the pinnacle of the industry, often referred to as the World Series of CTFs due to the pedigree of the competition's participants. Teams must qualify through preliminary events to compete. Other well-regarded competitions include the Plaid CTF, which is run by Carnegie Mellon's PPP team. Google has a CTF. There's a picoCTF designed for high school and college students, and many dozens of others are tracked on sites like CTFtime.org.

Whereas the Pwn2Own competitions are discovering original vulnerabilities, the CTFs are about discovering planted secrets. Both serve important roles for the industry. The CTFs are a legal sandbox for practicing techniques that would be illegal to use against real systems. And they've served as a recruiting pipeline. Top CTF performers are heavily recruited, or have historically been, I should say, because of what we'll see what's happening, heavily recruited by security firms, intelligence agencies, and tech companies. They build the shared culture and vocabulary of the field. And they often produce write-ups afterward, where teams publish detailed after-the-fact explanations of how they solved each challenge, creating a corpus of freely available security education.

Many of the researchers who discover major real-world vulnerabilities got their start, or stay sharp, through CTF competition.

Okay. So we have some sense for who Kabir is, and also for what CTF competitions have traditionally meant for the industry and to those who wish to use them to sharpen their hacking skills in a competitive legal framework and environment. And we know that Kabir has been one such person having participated in a number of CTF competitions and teams through the years.

So his blog posting is titled, sadly, "The CTF scene is dead." The details are very interesting, and many are very insightful and important. So here's what I want to share. He wrote: "Frontier AI has broken the open CTF format. The scoreboard does not measure human skill cleanly anymore, and the old game is not coming back. What makes me qualified to say this?" He says: "I started playing CTFs in 2021, the same year I started university. My first CTF was HCKSYD, a 48-hour solo CTF. I fully solved it and won in two hours."

Leo: Wow. Oh.

Steve: Yeah.

Leo: Wow.

Steve: And this makes me think, Leo, of your Capture the Flag format; right? I mean, it's very much - I mean of your Advent of Code.

Leo: Oh, yeah, it's like that.

Steve: Very much the - yeah. So he solved this 48-hour solo CTF in two hours. And he said: "I was completely hooked. That led me to win DownUnderCTF, Australia's largest CTF, with team Blitzkrieg multiple times. Blitzkrieg was one of Australia's strongest teams at the time. I later joined TheHackersCrew, an international top-tier team that was consistently ranked highly on CTFtime, the main global ranking and event calendar the scene uses as its scoreboard. With them, I competed in some of the most prestigious CTFs in the world, consistently placing well within the top 10 until the end of 2025.

"I am not saying this because I dislike CTFs. I'm saying it because CTFs were the thing that made me fall in love with security. They taught me how to learn, gave me a way to measure myself, and introduced me to many of the people I respect most in the field. Watching people pretend the format is still fine is frustrating because the old game is not there anymore." And Leo, if you think about online chess and cheating?

Leo: It's the same thing.

Steve: Same idea.

Leo: Same thing.

Steve: So he says: "What changed?" He says: "As AI tools ramped up in capability, especially when GPT-4 first came out, a significant percentage of medium-difficulty CTF challenges started becoming one-shot-able, meaning a single prompt from a user could produce the solve and flag. You could paste a cryptography challenge into ChatGPT, come back in 10 minutes, and have the solution. At the time, we didn't think too much of it. Hard challenges went mostly untouched, and the time save was not large enough to ruin the competition. The issue was never that AI could help. CTF players have always used tools. The issue is when the model does the reasoning, writes the solve, and leaves the human with nothing meaningful to do besides copy the flag.

"Enter Claude Opus 4.5: When Opus 4.5 dropped, the tone changed. Almost every medium difficulty challenge, and some hard challenges, became agent-solvable. Claude Code packaged everything into a CLI and made it easy to connect other CLI and MCP tools. It became trivial to build an orchestrator that used the CTFd API to spin up a Claude instance for every challenge. You could let the system run for the first hour, then only start working on whatever was left.

"That changed the game. Teams that refused to use AI were not just missing a convenience; they were playing a slower version of the competition. Open online CTFs started becoming a question of how quickly you could automate the easy and medium work, then how much human attention you had left for the hardest challenges. The scoreboard started measuring orchestration and willingness to use frontier models along with, and sometimes instead of, security skills. The effects were obvious. The CTFtime leaderboard started feeling wrong. Some legendary teams that were consistently near the top appeared less often. Player activity felt lower. Challenge developers who treated CTFs as an art form had less reason to spend weeks building something beautiful, if it was going to be beaten by an agent in minutes."

Leo: Yeah, why create a puzzle? Can I just show you, because I think I didn't really understand what CTFs were. So I asked my AI to give me an example of a beginner CTF. And this is a web page. The difficulty is very easy. And I think everybody in our audience will be able to do this. Find the flag. So you've given a web page. Now, here's the HTML code. And you can see hidden in a comment is the flag.

Steve: Ah.

Leo: Okay? So the question would be what's the flag? And it's pretty simple to solve. You'd right-click the page, view page source.

Steve: Right.

Leo: And you'd find the comment. So that's a very simple example of the kinds of challenges these are. Just to make it clear what we're looking at.

Steve: Right, right. So then, he says, on top of what I just shared, he said: "Then GPT-5.5 sealed the deal." He said: "I have been working heavily with GPT-5.5 and GPT-5.5 Pro after launch. By benchmark metrics, 5.5 is close to Claude Mythos capability, and Pro likely surpasses it. These models can one-shot Insane-level difficulty - active, leakless, heap pwn challenges on HackTheBox. They can solve a large portion of what a smaller CTF organizer can realistically produce. If you orchestrate Pro against Insane

challenges in a 48-hour CTF, there is a good chance you get the flag before the event ends. That makes open CTFs pay to win."

Leo: Oh.

Steve: Yes. "The more tokens you can throw at a competition, the faster you can burn down the board. Specialized cybersecurity models like alias1 by Alias Robotics are becoming less relevant compared to general frontier LLMs. The competition is turning into 'who can afford to run enough agents, with enough context, for long enough.'

"CTFs feel much more like a cheesable mess than a competition. Your performance in a CTF no longer defines your skill the way it used to. Recruiting security practitioners by CTF performance is becoming less meaningful. It's not even a particularly good measure of AI skill because most of the orchestration needed for CTFs is already open source or vibe codable."

So what's the "beginners are fine" take? "I have seen various takes that beginners can still learn from CTFs as they always have. These takes miss the scoreboard. CTFs were not just a set of puzzles. They were a ladder. Even as a beginner, you had something to climb. You could see yourself improve, solve more challenges, place higher, join better teams, and become more competitive over time. That feedback loop is breaking. If the visible scoreboard is dominated by teams using AI, a beginner is pushed toward using AI before they have built the instincts the AI is replacing. That is an anti-pattern. It prevents active learning, and active struggle is the bit that actually teaches you something. It's also completely demotivating to put in real effort and see no visible progress because the ladder above you has been automated.

"It also changes what challenges authors want to build. If beginner CTFs become another place where people quietly paste prompts and climb a scoreboard, authors have more reason to put their effort into learning platforms instead. At least on platforms like picoGym and HackTheBox, the expectation is education, and beginners are less incentivized to cheat themselves out of learning. Beginners are better off using picoGym, HackTheBox, and other lab environments where the point is actually learning, instead of pretending the public scoreboard still reflects human growth.

"So, 'CTF isn't dead'? I've seen some hopium posts" - love the word "hopium" - "some hopium posts about how CTF is not dead, it's just augmented by AI. They often point at CTFs like Defcon to argue that AI still cannot solve everything. That's true, but it's the wrong defense.

"The hardest top-tier finals have very few participants, and they're usually gated behind qualifiers that are easier than the finals themselves. If those qualifiers fall to" - meaning qualifying to compete in the final. "If those qualifiers fall to agents, fewer genuinely qualified people reach the challenges that still resist AI. A tiny number of elite finals does not save the open online format that most people actually play. The claim is not that every challenge is solved. The claim is that enough of the scoreboard has been automated that the scoreboard no longer means what it used to mean.

"What about the 'AI is useful for security research' take?" He says: "CTFs were never meant to be security research. They can showcase new and interesting techniques, but the CTF itself is not the point of discovery. Just because AI is useful within a field does not mean it belongs in the competitive landscape of that field. In CTFs, unrestricted AI removes the human from the puzzle almost entirely and reduces the art of security to a prompt. Sure, LLMs will keep getting better at security as long as CTFs are around, but that does not mean the competitive format is healthy. CTFs were an art form, a way to

share techniques with nerds, and a way to push the human bounds of security skill. That purpose is being stripped away.

"What about the 'LLMs are chess engines for cyber' take?" He says: "Chess has been dominated by computers for well over a decade. People use chess engines as an analogy for LLMs in CTFs, but they miss the point. Chess engines are not allowed during competitive play."

Leo: Right, right. That would take all the fun out of it.

Steve: Exactly. What's the point?

Leo: What's the point?

Steve: Yeah. "They're used for analysis, training, commentary, and practice. They enrich the game around the competition without replacing the person competing.

"Imagine giving every competitive chess player the best chess engine and letting them use it freely during matches. Would that be considered fair? Would it be fun to watch? Would it justify prize pools? Would it push the human limits of what could be achieved in chess? The same questions apply to CTFs.

"And organizers cannot fight back. CTF organizers have tried techniques to break or deter LLM solutions, but they are temporary friction at best. Claude Code does not meaningfully care about old refusal-string tricks anymore. Frontier models are getting better at noticing prompt injections. Web search capabilities weaken challenges based on technologies released after the training cutoff. Rules that ask people not to use LLMs are ignored and almost impossible to enforce in open online events. That leaves organizers in a bad position. If they make normal challenges, agents solve too much. If they make challenges deliberately hostile to agents, the challenges often become guess-y, over-engineered, or unpleasant for humans, too. That's not a real fix. It just makes CTFs worse for everyone."

Okay. And so what about "Just adapt, bro?" He says: "This take is infuriating. People I have always looked up to in the community have said it. Just adapt, bro. To me, it's completely nonsensical unless you explain what we're adapting into. If adaptation means building better tooling, CTF players already did that. If adaptation means writing harder challenges, organizers already tried that. If adaptation means accepting that the scoreboard is now an AI orchestration benchmark, then we should say that honestly, instead of pretending the old competition still exists.

"Even if organizers create guessier or more over-engineered challenges that current LLMs cannot solve, there are no good paths for players to learn the required skills while staying competitive. A few models from now, that point may be irrelevant anyway." Meaning it's going to keep getting better and better and kill that off. The trajectory of LLM security capability is moving too quickly for challenge design to stay ahead for long.

"So what's the aftermath? The scene that grew my love for CTFs is emptying out. The CTFtime leaderboard has almost no semblance of history or human skill anymore. The 2026 scoreboard is unrecognizable compared to every year before it. TheHackersCrew, alongside many other large and reputable teams, either do not play, play with far fewer people, or struggle to cut into the top 10. Unregulated cheating is through the roof. Some of the best CTFs, like Plaid CTF, are not even running anymore.

"These sentiments are not only mine. Many members of my local team, Emu Exploit, feel similarly. These are people who consistently attend the International Cybersecurity Championship, perform at the top level in bug bounty programs, compete in Pwn2Own, and present at conferences like Black Hat. The people losing interest are not casual observers. They are exactly the kind of people the scene used to produce and retain.

"The fun of CTFing is gone for many of the people who cared the most. The loss is not just a scoreboard. It's the ladder from beginner curiosity to elite competition. It's the craft of challenge design. It's the feeling that a clever human solved something difficult because they understood it deeply.

"That legacy is not being carried forward by online CTFs in their current form. The format is dead. Something else may replace it, but pretending nothing fundamental has changed only makes the loss harder to talk about honestly. It also gives AI shills more room to capitalize on the decline by selling mediocre wrappers back to the community that made the training data valuable in the first place.

"So what now? While a lot of what's happening in the CTF/AI space is super commercialized and out of our control, CTF has had a hugely positive impact on the industry. I have met so many kind, smart, and passionate people through CTFs. I have played some of the most beautifully crafted challenges and found some of the most intriguing unintended solutions.

"The community around CTFing has been an amazing place to learn, grow, and connect. That's something we should not lose, no matter where the competition goes. As a community, we should strive to stay together and build new avenues to stay passionate and keep learning. Security-adjacent social events like SecTalks, student conferences, and local meetups are great ways to stay connected and stay involved. Learning platforms and the communities they provide through platforms like Discord are also a valuable resource.

"While it may be a struggle to find an alternative to what we had, the amazing community we've built around it is more important now than ever as we find new ways to keep the competitive spirit alive."

So what's really happened here is that, sadly, the truth is AI killed something. And it's, you know, it's going to kill a bunch of jobs. That's true. And it's killed a competition where what it happens to be engineered to do is what humans were competing with each other for. So, you know, I think we all need to deeply understand, appreciate, and internalize the reality that entire field of software security as we have always known it has been forever changed this year. Mozilla knows this. Daniel Stenberg knows this. Kabir, who's mourning the death of his beloved and supremely valuable Capture the Flag training competitions knows this. Pwn2Own will die. The software bounty industry will dry up.

And bless their hearts, Zerodium, the firm that purchases zero-days for resale to undisclosed dark parties, will probably also blessedly die. All of the many various enterprises that have been built up over time, as a side-effect response to the fact that we've been able to easily create software systems that were too complex for us to fully understand, are likely to fall into the dustbin of history.

It is what it is. There's no point in mourning it. It's going to happen because now that, for the first time ever, we can have secure software, that's what we're going to have. What we are not going to have going forward is anything that exists solely because software has vulnerabilities.

Leo: It is, it's interesting. We've been talking, and I've been thinking about our Advent of Code code challenges. And already last year it was truncated. And I have a feeling this year it's not going to exist at all because it's just way too easy to one-shot it.

Steve: Yeah.

Leo: And I don't know if he could write them. He's very clever, the guy who does these. I don't know, though, if it's possible to write a coding challenge that can't be solved at this point by AI.

Steve: There are books of chess puzzles, right, where, like, there's a board with pieces on it, and it said, you know, Mate in 7 or something.

Leo: Oh, they're all easily solved, yeah.

Steve: Yeah. It's over.

Leo: Well, I mean, first of all, chess is computable. So it's done. It's solved. Every position. In fact, there are now...

Steve: And software is also computable.

Leo: Sure. It's deterministic. It's the same thing.

Steve: Yes. Yes.

Leo: Yeah.

Steve: Why would it be any different?

Leo: There are many, many more - there are many more variations in software. But yeah, I don't think it's any different. There are now books, there weren't when I was coming up as a chess player, but there are now books with thousands and thousands of positions, particularly in the end game. They're just solved. They're known. And it's like, if you get to this position, it's done. And, I mean, it's just changed everything in chess. But what it hasn't changed, and I'm actually relieved a little bit, is the fun and the challenge of playing head to head without an AI.

Steve: Yes. Yes. Yes.

Leo: You still can play a game. Even though you know if you had a machine on the other side you'd lose. And even if you're the best player in the world you would lose.

Steve: Well, and we know that you could take a picture of a crossword puzzle and drop it into AI, and it would fill it out.

Leo: Oh, easy. Yeah, that's actually somewhat spoiled Scrabble games because it's really trivial to cheat at Scrabble now. But here's the difference. You kind of can tell if somebody's cheating. Actually, it can even tell if you're cheating at chess because, even though that might be the best move, it's not a human move.

Steve: Ah, right.

Leo: So there's a slight, like, this is a little odd. I actually went back to the AI and said, okay, now give me a hard Capture the Flag one.

Steve: Oh.

Leo: And I thought this was kind of interesting. I suspect you would be able to solve this very quickly. Well, maybe not quickly. But I'll give you the harder example. I won't - I have the solution, and I even have some hints. But this is an example of what at least my AI thinks is a more difficult - and you could see the value of doing this, by the way. Let me see. I have to do this to get that to show up. There we go. It's called Reused Once. It's a crypto challenge. Hard.

Steve: Yup.

Leo: The theme is ECDSA nonce.

Steve: Elliptic curve.

Leo: Actually, that's a pretty big hint, nonce reuse.

Steve: Nonce reuse, yeah.

Leo: Forge an admin signature. So here's the login service signs user tokens using ECDSA over the [indiscernible].

Steve: Right.

Leo: The developers claim ECDSA is unbreakable. We use SHA-256. Good luck. However, you have intercepted two signed user tokens and noticed something odd. Here's token number one.

Steve: Very good.

Leo: And here's token number two. The signature on these well, the R value - I don't know what R and S are. But the R value is the same on both of them.

Steve: Yup. Yup.

Leo: Reused nonce. Uh-oh. So the goal in this is to forge a valid signature for this user if the verifier accepts...

Steve: Yeah, to actually do the math to determine what the key is because...

Leo: And this is like, if you got this on an SAT, you would plotz because the math is complicated. ECDSA signatures are computer does this, where D is the private key, K is the first signature nonce. Ah, here's the thing. We've got a repeated K so we can solve for Z, recover the private key, and forge an admin signature. You'd have to be pretty sharp to do this.

Steve: You have to really understand where this stuff comes from and how this stuff works. And why you can't reuse an elliptic curve DSA signature.

Leo: Very good, yeah. And so you've learned something by doing this. Of course you had to know something.

Steve: Oh, and that was Kabir's point is that, you know, struggling against these puzzles were incredibly valuable.

Leo: Yes. It teaches you. Yeah.

Steve: Yes. You had to reach out and extend and stretch yourself in order to understand what was going on. And AI killed it. AI killed chess. AI has now killed Capture the Flag. And the good news is AI's going to kill bugs in our software. So I would say it's worth it.

Leo: I think it was a fair trade. And you could still do Capture the Flag if you said you're not allowed to consult AI.

Steve: And that's the problem. In a global tournament, there are going to be weenies...

Leo: How do you stop people? Right.

Steve: ...who are going to cheat. You know?

Leo: It's exactly the same as chess.

Steve: Just like online chess, yes.

Leo: Online chess is eminently cheatable. In-person head to head chess, much more difficult. You have to have to some sort of vibrating device on your body. But not so hard when you're online. There's a lot of trust, basically, that you have to do when you play online.

Steve: Oh, it's known as sportsmanship. And unfortunately...

Leo: Yeah. And an understanding that the fun is playing the game, not having some machine solve it for you.

Steve: Yup.

Leo: Well...

Steve: And that's why I code in assembly language, by the way.

Leo: Yes.

Steve: The fun is in solving the problem.

Leo: It's in doing it. It's fun. And so Advent of Code will still be fun. The leaderboards are meaningless now because somebody will solve it in seconds using an AI. But if you are doing it with your friends, and you trust your friends not to do that, or...

Steve: And so, you know, that's really the way to fix the problem is remove the leaderboards, completely remove the notion of competition.

Leo: That's what he did. He did last year. He removed the leaderboards last year. You could still have a private leaderboard. We do. We have a TWiT leaderboard. And, you know, we know each other so there's a level of trust there. Nevertheless, you know, even last year there were some things - actually I found that AI could also be useful for learning. And there were some things that I had no idea how to solve, and I asked for kind of hints from AI or like, oh, I had read on Reddit there's a technique that will solve this, but I don't know what that technique is. Teach me how this works, the remove a single digit technique. And then you learn a new algorithm, and then you can apply it. And so AI can be used in a judicious way to help you learn, too. It's fun. It's really fun. I have to say.

So will we still do the Advent of Code, Paul and Darren and all the folks in TWiT? I don't know. We might. We might. Steve Gibson, you are a true sportsman, an assembly language coder. And as Anthropic's Claude tells me, you like to watch TV at night while you're sitting on your TV tray eating Chinese food and coding. Is that right? That's what Anthropic says. I don't know if that's right.

Steve: That's close.

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/