The Malicious Use of AI

Description: Was a U.S. law firm right to pay a $20 million ransom? Could Cisco have yet another SD-WAN zero-day in the wild? Why is it so difficult to author secure PHP code? Teens use "WeedHack" to spy and attack each other. Researchers create the first AI-enabled Internet worm. Google Chrome pops-up "Shop with confidence." What? The discovered and irresponsibly disclosed HTTP/2 Bomb. What Anthropic learns from their past year of Claude abuse: It's bad.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1082.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1082-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. This is a big day. Anthropic just released a new version of its AI, the Fable model. It's appropriate because we're going to talk about what Anthropic has learned from years of abuse of their AI models. We'll also talk about the malicious use of AI, some really scary examples. And which U.S. law firm paid a $20 million ransom to ransomware authors, and why? That and a whole lot more coming up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1082, recorded Tuesday, June 9th, 2026: "The Malicious Use of AI."

It's time for Security Now!, the show where we cover the latest security and privacy and a little bit of AI in here with this guy right here, Mr. Steve Gibson, the guy in charge at GRC.com. Hello, Steve.

Steve Gibson: Hello, my friend. Great to be - oh.

Leo: Got to do the "Live Long and Prosper" thing.

Steve: That's right. Great to be with you again.

Leo: So.

Steve: So for a long time we've been saying, predicting - it was a prediction, but not like - didn't take any great stretch of imagination that the bad guys would be using AI just like the good guys are.

Leo: Right.

Steve: And in fact the reason that Anthropic did its sort of semi-controversial Claude Mythos Preview limited, you know, strictly limited release was that their feeling was it was enough of an advance that, if the bad guys got a hold of it, then there wouldn't be time for the good guys to fix their broken code. So it turns out that there's a red team operating at Anthropic which has for the last year, from March 2025 to March 2026, been cataloging the abusers' use of their AI, various versions of Claude through the last year. And they've mapped it onto something we've never talked about before, which is the MITRE ATT&CK taxonomy. It's spelled ATT&CK because I guess you're going to have to be a hacker. And what they found is really interesting and just as worrisome as you could possibly imagine. I mean, like, you know, I'm not one to declare that the sky is falling. But it occurred to me.

So we've got Security Now! Episode 1082 for this June 9th, titled "The Malicious Use of AI," where we're going to, by the end of the podcast, have a bracing understanding of, like, the bad guys are not sitting around. They're not waiting. They're on this fast. And what an AI-enabled attack, like well-orchestrated malicious campaign can do is truly bone chilling. So, you know, I just hope that everybody who's got some jewels they need to protect are on the ball here and using all of the most state-of-the-art available tools. And one just dropped, like, what, an hour ago? A new update?

Leo: Yeah, Fable just came out, yeah. This is reputedly kind of a simpler stripped down Mythos.

Steve: Ah, okay.

Leo: And at the same time, apparently Anthropic turned on - because Anthony tried to put your show notes into the previous version 4.8. And it may have turned on some sort of gate that says - here, let me read you the actual text. It says...

Steve: I actually saw that.

Leo: Did you see that?

Steve: I actually saw that, yeah. And there was a slide switch where it would fall back to a less potent model if it thinks that you're asking for things that it's not sure it wants to give you.

Leo: It says it won't work on cybersecurity stuff.

Steve: Oh, oh.

Leo: And apparently your show notes are too dangerous. Opus 4.8, it said chats paused. Open 4.8 has safety measures that flag messages on most cybersecurity or biology topics.

Steve: Wait. 4.8?

Leo: They may flag safe normal content, as well. These measures let us bring you Mythos-level capability in other areas sooner.

Steve: Interesting. So because they're having a problem filtering, they did a crude filter. That is to say, you know, it's difficult. Like a perfect example is my show notes. There's nothing odd, you know, malicious in our show notes except we're talking about malicious things.

Leo: Right.

Steve: I mean, we're talking about cybersecurity stuff. And so the idea is like that's just, since they don't know that they can slice it correctly, they're just completely blacking it out. Nothing to do with cybersecurity, nothing to do with biology, because we don't yet know how to differentiate enough to give you access to that.

Leo: Now, I have just fed your show notes to the new model, which just came out this morning, called Fable. And Fable is kind of like Mythos; right? And it has no trouble, no trouble at all going through your show notes. No complaints whatsoever. So something's up. I did, in fact, and I think you saw it earlier on MacBreak Weekly, run some of my old Claude-generated code through Mythos saying find some security flaws. And it did. And it did a nice job. Stuff that it had previously audited and found flawless.

Steve: Saw no problems with.

Leo: Yeah. So I was very impressed, not merely with how quickly it worked and how well it did, but I was actually very impressed with the verbiage it used. It seemed quite impressive. And it's much faster. It whipped through a large number of files, both in Rust and Python and Go, and found faults, found flaws.

Steve: Now, I know, well, we know from the announcement that it uses - it's twice the token consumption rate as Opus; right? Fable does.

Leo: Yes. They say this right on the front.

Steve: And so it's twice as expensive, essentially.

Leo: Yeah.

Steve: Now, I've got the $20 a month plan. And I'm never hitting a ceiling because I'm not really, well, I have to say, though, that the only time I saw the thermometer start going up was when I gave it more of an agentic-y kind of thing to do, where it sat and churned for a while. And I thought, ooh, I wonder how expensive that was? And I went

over to check my account. It's like, oh, look, I just used up 20% of something. Whereas normally it doesn't even get off the ground for the little simple things I'm asking.

Leo: Yeah. So it was able to find these. I feel like this - it just feels a little smarter, a little quicker, a little more effective.

Steve: Wow.

Leo: I was quite impressed. This is all West Coast.

Steve: I'm just stunned by this pace, Leo. I mean, it's breathtaking.

Leo: Yeah. They just released 4.8, like, three weeks ago.

Steve: It's breathtaking.

Leo: It's crazy. Well, so I don't know if this is Mythos. But in a way something Mythos-like has arrived already. Which means I think you should start looking for tomorrow.

Steve: Time to start running your code through Fable.

Leo: Yup, yup. I did have it fix everything it found, by the way.

Steve: Very cool.

Leo: Yeah.

Steve: Okay. So in addition to getting to the malicious use of AI, where we're going to look at exactly what's going on, we're going to answer some questions. Was a U.S. law firm right to pay a $20 million ransom? Could Cisco have yet another SD-WAN zero-day in the wild? Like, really, Cisco? Really? Come on. Why is it so difficult to author secure PHP code? Turns out that teens are using something called WeedHack to spy and attack each other, which MacAfee's security people found and were quite disheartened to, like, see what was going on. Researchers have created the first AI-enabled Internet worm. And, oh, boy. The good thing is it's not clear that a worm makes anyone any money. And money is the name of the game for the bad guys now. Otherwise, it would be game over.

Also just a little editorial annoyance. Because while I was working I got a weird Chrome pop-up telling me that I could shop with confidence. I wasn't even using Chrome. I was in Firefox. It's like, what the heck? Where? What?

Leo: Oh. Oh.

Steve: We've also got - something was really wrong here - an irresponsible disclosure of a very bad problem that was discovered in HTTP/2. You know, we always had HTTP/1. Then we got 1.1. Recently we got, well, a couple of years ago we got 2. There's an HTTP/2 bomb which can basically bring any contemporary web servers to their knees. And the cretins who discovered it said, yeah, you know, what the heck? We're going to force everyone to upgrade by releasing it. Wow. And then we're going to get to what Anthropic has learned from their past year of monitoring Claude's abuse. And in two words, maybe it's three, I don't know if you count a contraction as two. Anyway, it's bad.

Leo: Wow. Well, we have lots to talk about and a Picture of the Week to come in just a little bit. You're watching Security Now! with Steve Gibson. And now, ladies and gentlemen, the Picture of the Week.

Steve: So there's no security angle here. But I just love this.

Leo: Okay.

Steve: I gave this the headline, "There may be hope for humanity after all."

Leo: All right. I'm going to scroll up. I haven't seen this before. All right. I'll let you describe this one.

Steve: So we have...

Leo: That's great.

Steve: We have two signs, the yellow diamond sign that says "Dip Ahead," right, like where there's just to warn drivers that there's going to be some sort of a dip in the road that they need to take advantage of. But then slightly after that along the road is one of those programmable boards for like whatever the people working on the road need to warn drivers about. In this case, the signage has been programmed to say, and this is again following the "Dip Ahead" sign, "Bring Chips." So...

Leo: To which our Discord chatroom has responded with this picture of you and me as, well, as chips. Let me pull up the image.

Steve: Oh, no.

Leo: In this case, "chips" is the California Highway Patrol, CHP.

Steve: Oh, goodness, yes.

Leo: Yeah, you look good in a uniform. Very nice. Thank you, Pretty For a CIS Guy.

Steve: Last time I was in a uniform was the Boy Scouts, and that was a - lots of stories came from that.

Leo: Oh, I bet.

Steve: Okay. So the large law firm Weil Gotshal and Manges, which reported, I mean, there are firms you never hear about. I've never heard about these guys. But $2 billion dollars in revenue last year. Right? So they're, like, so high end that they don't do any retail advertising, or they don't have any sort of a public presence at all. They're, whatever it is they've doing, you know, military, contractors, or who knows what, maybe international stuff. Anyway, they're raking in the bucks. They recently paid a $20 million ransom.

Leo: [Whistling]

Steve: Yeah. And I did the math. That's 1% of their annual take. So they did so in order to prevent the release of their confidential client data. And we don't know who their clients are. But again, $2 billion of revenue, and people we've never heard of before, they've probably got it going on. So the company said that their client's confidential data had been stolen from an external cloud storage site earlier this year by a group known as the Silent Ransom Group. The FBI sent out a private industry alert last year warning in advance that this Silent Ransom Group had been spotted and that they were specifically targeting U.S. law firms for their extortion campaigns.

Now, what I appreciate about this is the strategic value of targeting law firms for extortion. Everyone knows I'm not endorsing the practice, far from it. But I think that high-end law firms are an interesting and clever ransom target. We've been seeing and reporting on the surprising and welcome dramatic decline in the percentage of ransoms that are being paid lately. You know, many more companies are simply saying "no" now than were 10 years ago. Back then, "being hacked" was much more of a black mark on an enterprise's reputation than it has, frankly sadly, become today. You know, I'm not happy that "being hacked" is almost routine now. But with cyber attacks having become a nearly daily occurrence, no one who is observing them from a distance really cares that much anymore.

Now, as a consequence, companies are just saying no to ransom demands and putting out a press release saying that, oh, well, we were hacked, coming up with some spin about how the bad guys didn't really get anything of any super secret value, and then offering their customers 12 months of free credit reporting, as if that makes restitution. So nothing to see here. Move along.

Okay. So against that backdrop, rather than just stumbling upon targets of opportunity, the bad guys have needed to find targets where the "apologize, obfuscate, and move on" practice would not be available. And the confidential client data being retained by major deep-pocketed law firms just perfectly fits that bill.

There's no doubt that this Weil Gotshal & Manges knew that the disclosure of their clients' data would result not only in massive reputational harm, which a law firm can ill afford, but also in a mass of breach of fiduciary responsibility lawsuits, right, brought by their own current and past clients, whose data they have not protected adequately. We know that the FBI and others would have cautioned the law firm over and over that there's absolutely no guarantee whatsoever that the bad guys will honor their side of the

agreement - which, after all, is voluntary, and they're bad guys - by deleting all of the stolen data.

But given $2 billion dollars in annual revenue, even a $20 million dollar payout, you know, that's just a 1% tax on the company's annual revenue, when weighed against the alternative of absolutely certain disclosure of highly damaging data, well, that's a bet that's entirely understandable. As I've been observing, the only thing these criminals care about is money. They could not possibly care less about the actual data they've stolen from this Weil Gotshal & Manges. They know that the only chance they have of reliably obtaining voluntary ransom payments from their victims is if those victims believe from all past evidence that their payment of a ransom will result in the deletion, as far as they can tell, or certainly not the disclosure, of the stolen data so that it can then hopefully never be disclosed. Yes, it may be a bargain with the devil, but it almost certainly paid off.

So I think the takeaway here is the observation that the overall drop in ransom payment likelihood has predictably shifted the attacker's targeting to those specific enterprises, law firms being a perfect example of that, which have the most to lose if their stolen data is publicly disclosed. That's the threat. And so it's got to be a threat where the pain of that threat being actualized is so high that the company says, oh, well, we don't like it, but it's a gamble we're willing to take.

So companies like the regular, run-of-the-mill companies who merely have millions of small customer transactions, they're just increasingly shrugging off such breaches as unfortunate, but equally unfortunate nowadays, almost inevitable. So it's like, eh, we got it. We got hacked. How would you like a year of free credit monitoring? Well, we would rather that you hadn't been hacked. But, you know, thanks anyway.

If I were to share the news that another unpatched zero-day flaw in Cisco's SD-WAN manager was being actively exploited in the wild, our listeners could be forgiven for thinking that perhaps they were listening to a previous podcast. But sadly, no. In their reporting on this latest re-run of a story that pretty much writes itself at this point, you only need to change the date and tweak some CVE numbers, BleepingComputer reminded us.

They wrote: "Last month, Cisco" - now, so we had one, like, last week, right, a new SD-WAN zero-day flaw exploit. And to give some background, they finished their coverage, BleepingComputer did, saying: "Last month Cisco also tagged a maximum severity Catalyst SD-WAN Controller authentication bypass flaw (CVE-2026-20182) as an actively exploited zero-day to gain admin privileges on unpatched devices. While Cisco has not yet released patches for today's most recent problem, on May 14th it advised customers to upgrade to the software that had been fixed for CVE-2026-20182.

"In February, Cisco patched another Catalyst SD-WAN Manager information disclosure security flaw (CVE-2026-20133), which CISA flagged as actively exploited in late April and, two weeks later, warned that two more flaws (20128 and 20122) were being abused in the wild. In March, it also addressed and flagged a critical authentication-bypass vulnerability, which is, you know, the polite way of saying anybody who wants to can get in. And that would be 20127, that has been exploited in zero-day attacks since at least 2023."

And they said: "Over the past several years, CISA has tagged 90 Cisco vulnerabilities as abused in the wild, four of them in Catalyst SD-WAN Manager and six others exploited by ransomware operations." So 90 vulnerabilities in just the past several years abused in the wild. Cisco has certainly earned their reputation for providing hackers with a ready and, so far, unending supply of remotely exploitable security vulnerabilities.

Two months ago, on April 7th, so just over two months ago, today's the 9th of June, Cisco wrote: "For some time, we have been stress-testing our own products and infrastructure against the most advanced AI-powered security tools available, including Anthropic's latest unreleased AI model, Claude Mythos Preview. What we have found," they said, "has been illuminating. Now the real work begins. AI-powered analysis uncovers data at a scale and depth that legacy frameworks were not designed to accommodate." Whatever, okay, whatever that means. "This industry will recalibrate together, and Cisco is committed to leading that conversation."

All I can say is that I hope they mean it. I hope they really do suddenly care more than they have ever appeared to in the past. And given the evidence, it's like, how do you explain that? This is Cisco. You know, perhaps something needed to make security much easier for them to deliver, and perhaps "AI" will be that something that's been missing until now. It is inexplicable to me how a company that is so important and has been such a leader, you know, a pioneer on the Internet, could continue year after year to have so many damaging security problems. What is the culture over there? Recall that years ago they were surprised to discover that the firmware of their own machines had been shipping with embedded authentication credentials. In the firmware. So that anybody who knew the username and password could log in remotely.

Leo: I bet you Mythos would have found that.

Steve: Yes. Yes. Given access to the firmware, Mythos would have said the equivalent of WTF. So, but, you know, Cisco. Perhaps they are just really crappy at doing software, and the only reason, you know, the only reason they were ever on top is that they were first. And so it's like, you know, once upon a time they were the only game in town. And maybe they just always sort of sat back on their laurels and thought, well, you know, everyone's buying our stuff. It's broke, but what the hell.

Leo: I'm going to take a guess, a wild guess that part of it was the number of acquisitions they did. Because Cisco grew very fast by acquiring a lot of other companies.

Steve: That's fair.

Leo: And it'd be my guess that some of those companies themselves didn't have the best practices. And sometimes when you have mismatched systems you get these kinds of problems.

Steve: Well, perfect example was that Hilton attack. Remember?

Leo: Right.

Steve: It was they bought another chain.

Leo: Yes, it was Marriott, yeah, yeah, yeah. That's right, yeah.

Steve: And it was that, it was like they - but the thing they bought had some serious problems. And they just - at the time we argued that they didn't vet it as well as they should have. Which I think is a reasonable position to take. But, wow, Cisco, come on. You know, get your AI going and fix this because too much of the - now, of course the big problem, once there is firmware which the latest AI agrees has no more problems that it can detect, is how do you get it deployed? Because it's one thing to have it. It's another thing to have it out there running.

Okay. I've mentioned before that my desire to host web forums required me to run a PHP interpreter on a GRC domain, you know, forums.grc.com. And even grc.sc, that little link shortener, is also some PHP. But due to the long history of security incidents surrounding PHP, the idea of running a PHP interpreter on a GRC domain terrified me, and I was, and still am, unwilling to allow any such server to share a network with the rest of GRC's infrastructure. In other words, I took my own advice in the same way that I do for residential IoT devices, which is to firmly sequester those things whose security we have no control over and are inherently suspicious of on their own network. You know, contain them. Sandbox them.

Fortunately, my choice of the PHP-based XenForo for forums and the PHP-based NuevoMailer, which is what I use to send out GRC's weekly mailings. They've both been solid choices. And I've never had any problem with them. But I'm still not allowing anything that runs PHP anywhere near the rest of GRC's network. The reason I'm mentioning today this is that, once again, a PHP-based third-party Wordpress plugin has come under active, widespread global attack. And the means by which the plugin is being attacked is just so marvelously "PHP" that I wanted to take the time to share it.

Last Wednesday, the Wordfence Wordpress security company - and based on everything we've - we haven't talked about Wordfence for a long time. But they got a strong recommendation last time, and they'll get it again because I think anybody who is running WordPress stuff with any add-on plugins especially, which is where the problems generally are, WordPress, you know, itself has generally been so well cared for and maintained that we don't see problems in the core WordPress system. Anyway, they posted, Wordfence posted the news of this latest vulnerability which carries a CVSS of 9.8, which as we know is hard to achieve. You basically have to let anybody who wants to anywhere in the world crawl into your system and set up shop to get a 9.8.

Wordfence wrote: "On March 30th, 2026, we publicly disclosed a critical Remote Code Execution vulnerability in Everest Forms Pro, a WordPress plugin with an estimated 4,000 active installations. This vulnerability can be leveraged by unauthenticated attackers" - meaning anyone - "to execute arbitrary PHP code on the server, leading to complete site compromise. The vendor released the fully patched version on March 18th, 2026. Our records indicate that attackers started exploiting the issue on April 13th, 2026." So, okay, less than a month later. So March 18th, fully patched version that fixed the problem. April 13th it came under attack. So again, anybody who's keeping their site up to date, is checking for, like, checking for updates and following through with them, would have been safe.

"The Wordfence Firewall," they wrote, "has already blocked over 29,300 exploit attempts targeting this vulnerability." And they said: "Wordfence Premium, Wordfence Care, and Wordfence Response users received a firewall rule to protect against any exploits targeting this vulnerability on February 27th, way up in advance of the plugin being updated. Sites using the free version" - which is what, I mean, like why wouldn't you use the free version of this protection system of Wordfence - "received the same protection 30 days later on March 29th." So still well in advance of when the bad guys started attacking. They said: "Considering this vulnerability is being actively exploited, we urge users to ensure their sites are updated with the latest patched version of Everest Forms Pro, version 1.9.13 at the time of its writing, as soon as possible."

So as I said, we've covered the work of the Wordfence people in the past, and I have no problem allowing them to promote themselves by sharing their posting here since any site that has chosen to employ third-party Wordpress plugins would be well served to at least run the free version. And, you know, I'd pay something for the added protection if for no other reason to support them, in the same way that Leo and I do for Bitwarden.

Leo: Bitwarden, yeah.

Steve: Yeah. In other words, you know, these are good guys offering an important service at a reasonable price. And if you're running WordPress, and you've got, you know, random plugins that you've added on, you really ought to have Wordfence watching your back.

Then they explain what they found, which is really what I wanted to get to. They wrote: "Examining the Everest Forms Pro code reveals that the plugin uses the process_filter() function in the Process class to evaluate user-defined calculation formulas." Now, there's the key. User-defined, meaning visitor-based, calculation formulas. "The function concatenates submitted form field values into a PHP code string, which is then passed to the eval() function." Now, again, that phrase, that string of words should make anyone's blood run cold: "concatenates submitted form field values into a PHP code string, which is then passed to the eval() function."

So this is a variation of the infamous "Bobby Drop Tables" flaw. You know, any time any user-provided, you know, visitor, web visitor traffic input is passed to a function that might confuse data with commands, which is what PHP does in the same way that SQL can, that user-provided input must be scrupulously sanitized. I mean, really, you should never have a situation where that could be done. But if you have to for some reason, then really make sure there's no way that the user can provide something that can be switched from data to command, to prevent malicious users from managing to use a web form for their own command input.

Wordfence's write-up continues, saying: "Although user input is sanitized with sanitize_text_field(), this function does not escape single quotes" - and by escape they mean convert a single quote into something that isn't a single quote, but like carries the same meaning. It's known as "escaping" in programming parlance. "[It] does not escape single quotes or other characters that are significant in PHP code. For string-based fields (such as text, email, select, and radio fields), the submitted value is placed inside single quotes and directly added to a PHP code string. An unauthenticated attacker can exploit this by submitting a value containing a single quote, followed by malicious PHP code and a comment character, allowing them to break out of the string and inject PHP code that is later executed through the eval function.

"This makes it possible for unauthenticated attackers to execute arbitrary PHP code on the server by submitting a crafted value in any string-type form field, as long as the targeted form uses the 'Complex Calculation' feature. As with all remote code execution vulnerabilities, this could lead to complete site compromise through the creation of admin accounts, the use of web shells, and other techniques."

Okay. So this is exactly why PHP terrifies me. When I've made the mistake of stating that PHP is fundamentally insecure, our well-informed listeners have written somewhat indignantly to argue that it's entirely possible to write secure PHP systems. I assume that's true, since I have never had any trouble with XenForo, and their security record has been very good. Not perfect, but still very good. So I suppose a more balanced assertion on my part would be that authoring secure PHP websites inherently requires much more understanding of the security pitfalls inherent in the use of PHP, which are

many, than the typical PHP author possesses. In other words, you can write secure PHP, but PHP is targeted at people who don't.

Wordfence explains what their Wordpress application firewall has intercepted by writing: "The most common payload observed in our blocked requests attempts to create a new admin account named 'diksimarina' on the affected site. The attacker submits a value for a text field that begins with a single quote to close the wrapping string literal, followed by a PHP statement that calls wp_insert_user to create a new admin account with the username 'diksimarina.' The trailing // comment marker ensures the rest of the generated PHP code" - which was, you know, there in the original form - "including its closing quote, which the attacker put first, is treated as a comment and does not cause a syntax error because that would crash PHP, and then the attacker's code wouldn't get to execute.

"When the form is processed and the calculation is evaluated, the injected PHP code is executed, and the malicious administrator account is created. Once authenticated as a new administrator, the attacker can fully compromise the site by uploading web shells, modifying themes or plugins, or installing further backdoors to obtain persistent access." So the problem with PHP is that, while it correctly advertises itself as very easy to use, the less well appreciated fact is that it's also extremely easy to abuse. Thus it's running on a server at GRC, on its own network segment, with no contact to the rest of my stuff because I will never trust it.

Leo: I'm looking at the information about Claude Fable 5, which was released today by Anthropic, and Mythos 5. And I'm looking at the benchmarks provided by Anthropic. But man, they say this is even better than the Mythos Preview that they've been offering to some people. It is incredible. $10 per million tokens in, $50 per million tokens out. It is very expensive, although cheaper than Mythos Preview. They also say it can work longer than any previous Claude models. So they have a lot of benchmarks, a lot of examples.

Steve: I wish - I was thinking about this the other day. I wished that unused tokens built up a balance in your account.

Leo: I know.

Steve: Because my use is very erratic. And, you know, a lot of times I'm not using Claude for anything, and then I...

Leo: Well, do you have a subscription, or do you pay as you go? Because you can pay as you go.

Steve: I have Pro. I have a subscription.

Leo: Yeah, so the subscription's all-you-can-eat. And as you've probably noticed, if you really bang on it, it'll time out after five hours. It'll say, well, you've got to wait till 1:00 or whatever.

Steve: Right, right.

Leo: And you can also use up more tokens than you're supposed to in any given week or even month. But generally the all-you-can-eat is pretty good. The API is pay as you go. So if you don't use it at all, it's zero. So maybe - but the problem is, it's a lot more expensive to pay as you go than it is to buy a subscription for most people. I don't know what it's going to be like eventually. I think eventually Anthropic wants everybody to go the API route because I think for a lot of users, they're losing money on the all-you-can-eat. It's a buffet that some people are real pigs. But, you know, I'm just playing with it right now, and it's very fast. It also, they said, if it gets in a situation where you're asking about security stuff, it will fall back to 4.8. It's going to try to prevent you from using it to...

Steve: Yes. Yes.

Leo: And that's why we're getting that 4.8 warning is because...

Steve: Security or biology, apparently.

Leo: Yeah. I don't know about biology. But anyway, yeah. You can't make bugs, dangerous bugs with it either.

Steve: Yeah, exactly.

Leo: Of any kind, yeah. All right.

Steve: Okay. So McAfee's report, their headline caught my attention because it was "New Malware Targeting Minecraft Infects 2,000 Daily, and Teens are Becoming Attackers." So this is all pretty sad, but it's worth us knowing what's going on. McAfee writes: "McAfee Labs has discovered a massive, ongoing" - and massive because of how cheap it is - "ongoing malware campaign called WeedHack that disguises itself as free Minecraft mods and game clients to infect players' computers. Since January 2026, it has logged more than 116,000 victim infections, averaging between 2,000 and 3,000 new hits every single day. What makes WeedHack different from most malware is how cheap and easy it is to use. Typically, a hacker would pay hundreds of dollars per month to access attack tools through underground criminal networks." Now, this is all Malware-as-a-Service, which is the new thing.

They said: "WeedHack offers a free version to anyone with a Discord account. A premium upgrade, which includes the ability to secretly watch victims through their own webcam, starts at just $5 a month. This low barrier has attracted a younger crowd of would-be attackers. Many of them appear to be teenagers or young adults. Our researchers were startled to discover teens using these tools, not just for financial theft, but to harass and bully their peers, a pattern we've documented" - and like posting the webcam footage that they capture from other people's machines - "a pattern we've documented, and that makes this campaign especially concerning.

"WeedHack is a Malware-as-a-Service (MaaS) campaign, meaning it's a criminal business that sells hacking tools to customers, the same way a legitimate software company sells subscriptions. The 'product' in this case is malware that gets secretly installed on a victim's computer when they download what they think is a Minecraft mod or client. Once

installed, it can steal passwords, hijack accounts, and, for paying customers, it can give the attacker live access to the victim's screen, webcam, and files. The campaign operates a polished, professional-looking dashboard hosted openly on the Internet, not hidden on the dark web. That dashboard lets customers track their victims, download stolen data, and launch remote access features, all from their browser.

"One of the most disturbing findings from our investigation is how WeedHack is being used. While monitoring the campaign's Telegram channel, which had over 850 members during the time of our research, we observed that many customers appear to be teenagers and young adults, and a significant portion were using the remote access tools not for financial gain, but to harass and intimidate other players. We observed attackers recording victims through their webcams without consent and sharing those recordings in a Telegram channel as trophies. Others used knowledge of victims' IP addresses and system access to threaten them.

"It's important to note that, at the current time of publishing, the Telegram channel has been taken down, and no replacement channel has appeared. McAfee is continuing to monitor any new channels that may be established by the threat actors for further communication. Still, what we observed is a form of cyber bullying with unusually invasive tools behind it. If you or your child has been contacted by someone online claiming they've hacked your computer, have your webcam footage, or know your IP address, take it seriously. Do not follow the attacker's instructions, it only makes things worse. Tell a trusted adult immediately (parent, guardian, school counselor). Contact your local law enforcement. This may constitute criminal conduct. And do not engage with the attacker or attempt to negotiate.

"So, how Do People Get Infected? WeedHack spreads in two main ways, and the campaign even provides its customers with step-by-step tutorials on how to carry out both. First, fake YouTube videos: Attackers create convincing YouTube videos reviewing or demonstrating Minecraft clients and mods. The videos are well-produced, some include voiceover narration, the link to malicious download sites in the description and comments is present. One video McAfee identified had over 7,500 views before being flagged. Comments are also sometimes planted by the attackers claiming the files are safe.

"Second way, fake mod websites: WeedHack instructs customers to build convincing-looking websites that mimic official Minecraft mod pages. These sites are deliberately designed to show up high in search engine results for popular mod names, a tactic called SEO poisoning. Some fake sites include fake security warnings, Discord links, and GitHub references to appear legitimate. In one case, a site warned players to 'only download from us,' while actively distributing malware. Minecraft clients and mods specifically targeted include: Meteor Client, Radium Client, Wurst Client, LiquidBounce, Impact Client, Future Client, and others.

"So what happens when you're infected? Infection occurs in four stages that happen silently in the background after a victim opens the downloaded file. First stage, first contact. The malicious file launches quietly (without showing a console window), connects to a hidden network, and phones home to receive further instructions. It uses a sophisticated technique involving the Ethereum blockchain to locate its command server in a way that's difficult to block or take down." Remember we talked about one such method using DNS domain names, which are created dynamically based on a timestamp. This uses the Ethereum blockchain.

"Stage two, taking hold. The malware disables Windows Defender protections, gathers detailed information about the victim's computer (processor, graphics card, RAM, operating system and so forth), and takes a screenshot of their screen. It then steals their Discord tokens, browser passwords, and cookies. Stage 3, digging in. The malware installs itself so that it automatically restarts every time the victim logs back into their

computer. It sets up a hidden scheduled task that runs continuously with the highest system privileges."

And finally, "Stage four, obtaining full access: For premium customers, an additional component is installed that connects the attacker to the victim's computer in real time. This includes live screen sharing with keyboard and mouse control, webcam access, keylogging (recording every keystroke), a reverse shell (full command-line access to the computer), and the ability to upload or download any files. A separate component specifically hunts for Telegram credentials and cryptocurrency wallets, sending that data to a different server every five minutes.

"What Can Attackers Steal? The Free tier supports the theft of Minecraft session IDs, which are used to hijack Minecraft accounts; saved passwords and cookies from 36 different browsers; credentials from Discord, Steam, and Telegram; browser-based crypto wallets (56 are currently supported) and desktop crypto wallets (12 are currently supported); files matching 24 different search keywords; screenshots of the victim's screen; and system information (computer name, IP address, hardware specs). Then for $5 a month, which is the premium tier, you get live webcam access in addition to all those things, live webcam access; live screen sharing with keyboard and mouse control; keylogging (every key the victim types); full remote shell (command-line control of the computer); and file management (upload, download, and delete files remotely)."

Okay. So just to be completely clear, what now exists is a service which, for as little as $5 per month, apparently in a play for usage volume, anyone, and often teens, no longer need to have any - not any - hacking skills. Apparently all they need is some marketing skills. All the hacking, all the technology, all of that has been done for them. They're able to subscribe to this new Malware-as-a-Service, WeedHack, out on the public open web and then trick others, using their marketing skills, into downloading a Minecraft mod or client that then gives them access to that infected user's saved passwords and cookies, their social media credentials, their crypto wallets, and more, including their webcam and full remote keyboard and mouse access to their computer.

And the service is going gangbusters, logging between 2,000 and 3,000 newly infected victims per day, with more than 116,000 victim infections spotted since this past January. So, Leo, the world we live in today. Wow.

Leo: Yeah, no kidding.

Steve: You know, basically turning teens into criminals. Because this is all criminal abuse. I mean, this is criminal network intrusion thanks to a third party that's only asking for $5 per month and taking all of the hard work, all of the knowledge, all of the technology out of the loop. You pay them $5 a month.

Leo: What's worse is it's using kids to get to their - I presume to get to their parents' accounts; right? I mean, because...

Steve: Can be. But...

Leo: Yeah, I guess the kids might have money. I don't know.

Steve: Yeah, it's not clear where - I guess the kid said, hey, you know...

Leo: Some adults play Minecraft, but it's mostly kids; right?

Steve: It's got to be, "Hey, Mom, can I charge $5 a month for this cool service that I found that will allow me to do something with Minecraft." And Mom says, you know, okay, fine, whatever.

Leo: Yeah, happens all the time, yeah. Wow.

Steve: Wow.

Leo: It's so evil. It's so bad.

Steve: It is. It is. And teens are like, well, wait, it's a service on the Internet. What do you mean I'm breaking the law? What do you mean I'm a criminal?

Leo: It's actually smart because, yeah, exactly. This is taking advantage of their naivete.

Steve: Yup. And getting $5 a month out of all of their parents' credit cards.

Leo: Right.

Steve: Okay. So what have we done? You know somebody was bound to try it. So far, it's been contained inside a lab. Yikes.

Leo: Uh-oh. I don't like that. I don't like that open.

Steve: No. Yes. We've known of other things that were supposed to be contained inside a lab and got loose. Researchers with the University of Toronto and the Vector Institute wondered what a contemporary AI-powered network worm might look like, and how effective it might be. So of course they made one. The paper they just published is titled "AI Agents Enable Adaptive Computer Worms." And they explain: "In our pursuit of new knowledge" - that's always the excuse, right, or the justification. "In our pursuit of new knowledge to enhance the security of artificial intelligence, we uncovered a cybersecurity threat with implications across society."

Okay. So, you know, since the idea of an AI-enhanced network worm is not a stretch for anyone, I'm just going to share the high points from the research overview which they published. However, even this overview makes me somewhat queasy. Here's what they wrote. They said: "Large language models now demonstrate the capacity for structured problem-solving which, combined with tool access, enables agentic AI systems to solve complex tasks. We show that when these capabilities are embedded in a self-replicating agent, they produce a fundamentally new cybersecurity threat: an adaptive computer worm that devises target-specific attack strategies to gain control of machines and

spread across networks. Each compromised machine becomes part of the worm's own infrastructure, providing compute or reach for further attacks.

"A computer worm," they write, "is self-replicating malware that spreads across a network without human intervention. The WannaCry worm in 2017 disrupted critical infrastructure across 150 countries by exploiting a single vulnerability. Traditional worms can be stopped by patching the specific vulnerability they exploit. Our adaptive worm cannot be stopped this way: it uses a recursive reasoning loop to detect and exploit diverse vulnerabilities as it propagates.

"We demonstrate these capabilities in a controlled experiment: a prototype AI-driven worm powered by an open-weight LLM running locally, propagated across a heterogeneous network of Linux, Windows, and IoT devices with common corporate network vulnerabilities. The experiment was conducted in an isolated virtual network. We believe this work highlights three important dimensions of the impact of AI on the cyberthreat landscape.

"First: It establishes a qualitative shift in threat capability. The worm replaces fixed exploitation code with goal-directed reasoning that adapts to the vulnerabilities of each encountered target in real time. Our agent self-replicates across networked devices, subverts control of systems, and self-sustains on stolen resources.

"Second: The AI-driven worm requires only an open-weight model that can run on a single, local GPU. It does not rely on any commercial AI platform. This renders vendors' centralized safety controls, including service refusal, content filtering, and rate limits structurally irrelevant. The worm's tiered design, where each compromised GPU-equipped node provides reasoning for lightweight agents on downstream devices, extends the attack surface to any networked device." And I'll note that it gets smarter as it propagates; right? Because it's continuing to have access to all the GPUs it's already taken over. So that's kind of creepy.

"And, finally: The traditional economic barrier in cybersecurity collapses." The traditional economic barrier in cybersecurity collapses. "The worm parasitically uses the victims' own computational resources, reducing the attacker's marginal cost to zero. As consumer devices increasingly support LLM inference" - meaning they're getting the GPU compute locally, the reasoning resources available to such adversaries grow accordingly.

"This work provides empirical evidence that autonomous cyber offense has crossed from theoretical risk to demonstrated capability, a challenge that spans AI research, cybersecurity, and public policy. We believe this transition demands rigorous, transparent evaluation of model capabilities across the open and closed-weight model ecosystems." To which I say, yeah, good luck, because there's going to be unrestricted open weight models, I mean, there already are. They're only going to get better. So I'm not sure what that conclusion is supposed to mean. You know, we believe this transition demands rigorous transparent evaluation of model capabilities across the open- and closed-weight model ecosystem.

What this actually means, you know, and for the first time demonstrates, is that the defenders of cyberspace had better get serious about tightening up their code. This leaves the huge problem of the existing installed base of systems. And that's where we're really going to have our next problem. We know that. Because they're not going to update themselves. And, you know, many will never be updated without some form of more than typical intervention. I don't know how this happens. But it is the big problem.

The one bright spot here is that, you know, knock on wood, we seem to be past, as I mentioned earlier, we're past those frolicking days of uncontrolled Internet worms. Most mischief now is about bad guys focusing solely upon making money. And Internet worms

do not do that, you know, the way that targeted extortion can. The place I can see a worm being deployed as an offensive cyber weapon is not by a criminal organization that wants to make money as its first priority, but by a nation-state like the U.S., China, North Korea, or Russia. And in that case, its worm code would be carefully written to restrict its reach and spread to within a targeted geography.

So this feels more like an interesting academic exercise. This is not to say that someone might not release such a thing just to see what it could do. That's always a possibility. But as these researchers also noted, the world does not yet have sufficient potential victims with inference engines capable of supporting a roaming large language AI model. By the time that changes, the world's exposed vulnerabilities should be, you know, so new that any potential worm would starve. So, you know, hopefully we're going to get the Internet cleaned up. And as we evolve to next-generation systems that do have the required inference engines, they'll be running newer firmware or newer software, hopefully that will never have the vulnerabilities that the current thankfully dumb installed base of hardware does. Wow.

And Leo, I mentioned to you, I think before we began recording, that the weirdest thing happened. Or I guess I did run through it in the things I want to talk about. As I was writing these show notes, I was rudely interrupted by an unsolicited Windows 10 notification from Google Chrome. I wasn't using Chrome. As we know, I don't use Chrome unless there's no alternative. Sometimes there's a site that Firefox won't display. OpenTable.com is the one that keeps biting me, and now I've learned, I just have to go over and use OpenTable from Chrome. I haven't even used Chrome in recent memory. Yet what we know is that today's web browsers are all running agents in the background which serve to keep those browsers up to date. And that's a good thing. We know that we want our web browsers patching themselves to keep themselves current. I'm all for that. But the operative phrase here is "in the background."

When Chrome, which I have installed, but as I said I've not been using without, you know, like for any reason like for quite a while, certainly since that machine was booted, when it without invitation pops up a notification telling me that I can "Shop with confidence" and that I can "Track prices across the web and get alerts if the price drops on any site," first of all, I don't do that. But it's no longer "in the background," and it's an annoyance in my foreground. The Internet appears to be silent on this issue. I went searching, like is this happening to people? It's like, what? So I don't know what's going on. Perhaps this is Google trying to get traction for some of its new agentic AI crap.

In any event, I hope more than ever that Mozilla is able to somehow keep Firefox alive. A web browser, you know, is serious business, and keeping one going and secure and up to date with the never-ending and ever-changing World Wide Web Consortium standards, it takes a huge amount of work. So I appreciate what Mozilla is doing. But I need my Firefox. And I sure want as little of Chrome as I can get.

Leo: Yeah, I agree, yeah.

Steve: And I put, I grabbed a copy of this notification from Chrome. I put it in the show notes. It's like, what? I don't want to hear from a browser that I'm not using about shopping tips. Thank you [crosstalk]...

Leo: [Crosstalk] Windows doing this. Right? How could Chrome - well, you have notifications turned on in Chrome, I guess.

Steve: Yeah. And they are off now. But I hadn't turned them off before. So get out of my way, Chrome.

Leo: Yeah.

Steve: Okay. After our next sponsor note, I'm going to talk about this HTTP/2 Bomb attack and how annoyed I am with the people who just said, well, we're going to make everybody update their web servers by publishing an exploit. What?

Leo: Huh? Well, that'll work.

Steve: It's 2026, guys.

Leo: All right, Mr. Gibson. What else have you got for us?

Steve: Okay. So it's been a while since the abuse of a core Internet protocol was able to take down a wide variety of servers. But the recently discovered and, as I've said, very irresponsibly disclosed HTTP/2 Bomb attack, as it's called, can knock down NGINX, Apache, IIS, Envoy, Cloudflare's Pingora, and presumably any other modern web server that accepts and terminates HTTP/2 connections and queries, as all of the current state-of-the-art web servers do.

An independent observer of this wrote: "Since the bug is an HTTP/2 protocol bug, other services may also be affected. With NGINX (which is used in hardware load balancers, the most popular open-source reverse proxy, and ingress controller for Kubernetes), Envoy (which is a reverse proxy for large cloud and tech companies such as Google, Amazon, Netflix, and Airbnb), and Azure, which uses IIS, all of those can be knocked off the air with this thing." So they said: "A large portion of modern public-facing web infrastructure is affected."

And when they say "affected," they're not kidding. They wrote: "Given these circumstances, we had to take a close look at the write-up and assess the impact in order to determine which of our customers were affected and to incorporate the new attack vector directly into our platform." And they wrote: "And, oh boy, that attack is effective. With one single notebook we were able to bring down ANY of our own HTTP/2 servers, small and large.

"A single attacker can consume 20 to 30GB of RAM on the target." This is a resource consumption attack where 32GB are consumed and then locked, and it makes the server crash. They said: "RAM/memory remains locked even after the attack is stopped; this allows for a 'low and slow' attack, in which the attacker starts with a low connections and stream rate, but gradually consumes more and more of the target's RAM resources over time. Once a certain RAM usage threshold is reached, the affected NGINX instance crashes and must be restarted via a hard reboot. So this creates permanent damage."

And they said: "With these attacks, even a small botnet of just 10 bots can take down services of any size. For a low-and-slow attack with fewer than 10 requests per second, a botnet of only 100 bots is sufficient. And due to the slow query rate, such an attack would be undetectable and unstoppable by any web application firewall, which would normally be blocking high rate attacks."

Okay. So given the discovery of a truly devastating attack against pretty much all of the current Internet infrastructure, the fact that its discoverer chose to release the details without coordinating with the rest of the industry, well, in this day and age it's truly unforgivable. We first encountered "Calif," that's the name of this group, C-A-L-I-F, like as in abbreviation for California, but I don't know what it is, Calif. We encountered them recently, and I didn't think much of them at the time. Now, unfortunately, I do think something of them, and it's not good. Their website's home page declares in large font type "Pushing the frontier of vulnerability research with AI," and the sub-head is "Leet hackers and top models, what could go wrong? Let's find out." Wow. Okay.

Their posting about this sets the tone for them and for their site. One week ago today, on June 2nd, their blog posting carried the title "Codex" - you know, OpenAI's code AI - "Codex Discovered a Hidden HTTP/2 Bomb." And they wrote: "14 years ago" - so this is written in the first person by the person at Calif who was the discoverer. So he wrote, or she, they: "14 years ago I helped break HTTP header compression, then was asked to review the fix, which became part of HTTP/2. Life has come full circle: Today we're releasing an attack I missed. We're publishing HTTP/2 Bomb, a remote denial-of-service exploit against most major web servers, including NGINX, Apache httpd, Microsoft IIS, Envoy, and Cloudflare Pingora. The vulnerable behavior exists in each server's default HTTP/2 configuration.

"The attack was discovered by OpenAI's Codex, which chained two techniques known to humans for a decade: a compression bomb and a Slowloris-style hold. The bomb targets HPACK, HTTP/2's header compression scheme: one byte on the wire becomes one full header allocation on the server, repeated thousands of times per request. The hold is a zero-byte flow-control window that keeps the server from ever freeing any of it.

"A curious search on Shodan revealed more than 880,000 websites supporting HTTP/2 and running one of these servers." In other words, before releasing this, or at the time of its release, they know, thanks to Shodan, that 880,000 websites can be brought down with this. Yet release it they do. They said: "Though many" - oh, here, so here's a caveat to that. "Though many sit behind a CDN, which is much harder to bring down." They wrote: "A home computer on a 100Mb connection can render a vulnerable server inaccessible in seconds. Against Apache httpd and Envoy, a single client can consume and hold 32GB of server memory in roughly 20 seconds."

They then get into the details of this potentially debilitating vulnerability that's sufficient for anyone proficient to design an attack. But that won't be necessary because not only did these irresponsible jerks describe something for which they knew there was no current defense, but they also published a fully working proof-of-concept exploit. Not surprisingly, the folks over at Envoy, who produce that reverse proxy front-end used by companies such as Google, Amazon, Netflix, and Airbnb, were not humored by the behavior of these irresponsible glory hounds, so they posted to the feedback thread for this announcement blog posting.

They wrote: "OP ignored responsible disclosure policy and released a zero-day for Envoy's ecosystem. Envoy community was in process of releasing a patch for this problem." And then they have a link to Envoy security advisories, which is at GitHub.com.

To which, that posting, the Calif guy replied: "Thanks for fixing the issue so quickly. This is a win for Envoy users." Yes, you jerk. "We believe the traditional disclosure model is increasingly outdated in the era of AI-assisted vulnerability discovery, and we explain our rationale for disclosure in the post." So they've unilaterally decided, well, that old responsible disclosure doesn't make any sense anymore. So we're going to, you know, stir this all up.

The Envoy guy replies: "And irresponsible disclosure is a huge loss" - oh, so this guy says this is a win for Envoy users; right? It's a win for Envoy users. Envoy replies: "And irresponsible disclosure is a huge loss for Envoy ecosystem and possibly wider industry. Did you disclose this to all H/2 implementations? This should have really been coordinated via VINCE to make sure all H/2 vendors are aware. And if the 90-day disclosure policy is outdated, what is the new policy that you believe is appropriate? You have filed advisory on May 27th and published this blog on June 2nd. So is your new embargo policy four days?"

Calif responds and finishes with: "We disclose details once we believe that anyone monitoring public commits could reproduce the issue using AI-assisted analysis." I guess that means instantaneously. "In our view, withholding information after the relevant commits are public does more harm than good. We recognize that reasonable people may disagree, and we respect that perspective." Well, thanks a lot for your respect. What a mealy-mouthed position.

I suppose we're going to be seeing more of this sort of thing as those who could not have discovered this attack on their own now use AI to find the attacks for them. In the era before AI, acquiring true expertise would generally be accompanied by the acquisition of some maturity about it. Or they would value the attack they discovered because it was so hard to discover it, and so then they would responsibly disclose it to the people who it affected. Now discovering new attacks is free. They don't cost anything.

When AI hands someone who has never done the hard work, they don't have the maturity to guide their handling of such a gift. In this case it is utterly unconscionable that this exploit would have been publicly posted without a far more widely coordinated vulnerability and private vulnerability disclosure. NGINX has scrambled to assemble a patch and has made it available through standard update channels. But we know that's not the same as it being online. Apache's fix exists in a standalone module that has not yet been bundled into any release that package managers will pick up. So again, not enough time. Microsoft IIS has no patch, and no CVE has even been assigned to the IIS variant yet. And we know that Envoy also has just had to scramble.

As I said, I suppose we're going to be seeing more of this sort of thing in the future. I'm not unhappy that I'm still running an HTTP/1.1-only web server for a change. I'm sure that by the time I'm ready to deploy GRC's new servers, which I've talked about recently, Microsoft will have updated IIS to protect from this. But in the meantime, what a mess for everybody else. Wow.

Another little related AI note from me. I've mentioned that I am currently working, speaking of GRC servers, to reduce the purchase friction for GRC's software by supporting a range of one-click purchase options such as PayPal, Google Pay, Apple Pay, Venmo and so forth. Since my plan is to create a few more low-cost commercial products before I plow back into SpinRite for Windows, I want to make purchasing those as simple as possible. So I've been working to upgrade the eCommerce system which I wrote 22 years ago. I briefly flirted with the eCommerce provider Stripe since I liked their integration solutions. But I decided to go in another direction.

During that brief flirtation, they got hold of my email address, actually an alias that I used just for them, but still, it's alive. So I've been receiving occasional notes from Chet at Stripe, reminding me how wonderful they are. And they just kind of come in, and I ignore them. I've noticed that every email from Chet contains a link inviting me to set up a time for further discussions, you know, of the wonders of using Stripe. But I've just been ignoring those emails and letting them go unanswered. Until this past Sunday evening, when I decided to explain to Chet that since they do not support PayPal payments, Stripe is a nonstarter for me.

Yesterday morning a reply was waiting for me from my Sunday evening email, informing me of the good news that Stripe did support PayPal payments so that, as the email put it, it didn't need to be one or the other. And as with every email, it quickly ended with a link to "book a time" for a Stripe discovery call. Now, what I'd failed to mention to this Chet was that I needed to have everyone able to use PayPal, including U.S. domestic purchasers. And that is not something PayPal allows anyone else to do. So Stripe does offer PayPal, but only for some international users.

After jotting a short note back to Chet, I was hit by the question, am I interacting with a person? Because thinking back on all those previous emails and its response to me when I finally did answer it, I suddenly had that question. Is there actually any Chet at the other end of this email dialogue? I realize that these days it's entirely possible now that all of this sort of frontend sales lead development has already been automated by AI. I'm being pushed by automation to click a link to make an appointment with a real person. And the cost of that pushing from the pusher's end may have been reduced to zero. They don't need to be paying a human any longer. It's not a great job, you know, even when they were paying a human. But still, now all of the cost of that interaction is at my end. It's actually a new form of business spam.

And the other thing that clicked into place for me is that it's becoming prevalent because one of the things I've been noticing is the degree to which an increasing percentage of other enterprises that I'm noticing are having stuff outsourced. When I was interacting with DigiCert a few months ago, I noted that many of the links which looked like theirs actually pointed to Salesforce.com. DigiCert is outsourcing a large chunk of their customer service communication handling.

Now, a second-order consequence of this, of this increasingly prevalent outsourcing, is the degree I guess of what I call "presence broadcasting" has been steadily increasing. I've been noticing it happening. In the good old days, when a company needed to design and develop their forward, you know, their outward-facing communications for themselves, everything was bespoke. It was varied, and it was minimal, being only what they really needed for them, what really made sense for them. But now when a company signs up, for example, with Salesforce, they simply check off all of the various crappy outreach services they want to offer in their name and which inherently subjects everyone to that they can find. And that service provider then makes it happen.

And today, now we add to that a patient, never tiring, proactive emailing AI agent which is going to have zero-cost conversations as a means of what used to be working the phones, or in this case working the email. So it seems clear to me that many businesses are soon to become much more annoying. I just, you know, I had this weird thought, Leo, like as I replied to these emails, they're all very succinct, very short, from Chet. And they all invite me to, you know, find a time when we can have a conversation. And I doubt that there's any Chet at that end because nothing I've done really requires one.

Leo: Yeah.

Steve: And I don't know if you've noticed, but I'm feeling a homogeneity among very different companies that have identical-feeling outreach, and I realize it's because, oh, they didn't write it. They're now subscribing to an outsourcer that just does this. And unfortunately, it just means we're going to get a lot more of this crap.

Leo: As Internet users we get good at ignoring stuff. Pop-ups, spam, I mean, it's just endless.

Steve: It is unfortunate.

Leo: Yeah, yeah.

Steve: Okay. We have two more breaks. Let's take one now, and then we're going to get into looking at Malicious Use of AI. And we'll take our next break, you know...

Leo: Okay.

Steve: ...about halfway through that.

Leo: Back to you, Steve.

Steve: And it is rapidly getting scarier.

Leo: It is.

Steve: So, okay. We all knew it was coming, but it is no longer "coming." It has arrived.

Leo: It's here.

Steve: It's here. Last Wednesday, Anthropic published a red team report which examined the detected abuse of their Claude AI by malicious actors. We need to understand and examine how AI is being used by those who are, you know, aiming at nefarious ends in order to protect themselves.

So the report's three authors open their report by writing: "We've spent the last year investigating how threat actors are weaponizing AI to conduct cyber operations. Today we're sharing a new analysis that maps these real-world attacks onto the MITRE ATT&CK framework, a database of tactics and techniques used by cyberattackers. Doing so reveals patterns that challenge traditional assumptions about cybersecurity, for example, that the level of risk a threat actor poses can be assessed via metrics like technical sophistication or breadth of techniques. We partnered with Verizon to include some of these results in their 2026 Verizon Data Breach Investigation Report, and are publishing this report to offer a longer-form analysis of trends we see in AI-enabled cyber operations."

Okay. So what I'm about to share from their report, you'll hear these researchers referring to "accounts." The "accounts" Anthropic is referring to are Claude AI accounts whose holders were attempting to use or abuse their access to Claude AI for malicious purposes.

The other key to understand is this MITRE ATT&CK framework, which we've never had the occasion to look at closely. The MITRE ATT&CK homepage explains. They said: "MITRE ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations." And I'll just note that at the end of this report they observe that the MITRE ATT&CK knowledge base is going to need updating

based on the impact of AI. So MITRE says: "The attack knowledge base is used as a foundation for the development of specific threat models and methodologies in the private sector, in government, and in the cybersecurity product and service community. With the creation of ATT&CK, MITRE is fulfilling its mission to solve problems for a safer world by bringing communities together to develop more effective cybersecurity. ATT&CK is open and available to any person or organization for use at no charge."

Okay. So this MITRE ATT&CK database is really nothing more than a well-thought-out and carefully constructed taxonomy of all the various things bad actors have been seen to do through the years. And, right, it makes sense to have like a common vocabulary, a common enumeration system, where we can say, like, this technique and this tactic were used, and have those meanings well defined and described.

So, for example, it breaks malicious conduct down into 15 categories: reconnaissance, resource development, initial access, execution, persistence, privilege escalation, stealth, defense impairment, credential access, discovery, lateral movement, collection, command and control, exfiltration, and impact. And so basically those 15 broad categories, they're enough to contain whatever we see. And then each of those broad categories of malicious conduct is then broken down into a specific behavior.

I'll just give you an example of one. So, for example, taking that first category which was reconnaissance, that's broken down by this MITRE ATT&CK framework into 12 specific techniques of reconnaissance: active scanning, gather victim host information, gather victim identity information, gather victim network information, gather victim org information, phishing for information, query public AI services, search closed sources, search open technical databases, search open websites and domains, search threat vendor data, and search victim-owned websites. So again, this is meant to be a comprehensive description of anything that the bad guys do.

And so what the Anthropic researchers have done is that they took everything that they saw during a 12-month period, from March of '25 to this past March of '26, and plugged all of the behavior into this MITRE ATT&CK database in order to talk about it. And so the terms from the database are what I will be then describing. It is a widely agreed-upon system for categorizing and naming.

So here's what we know from their examination. The three researchers write: "For this study, we analyzed 832 accounts" - again, accounts meaning bad guys had a Claude AI account which is what they found, where they found the misbehavior. So "For this study, we analyzed 832 accounts associated with malicious cyber activity over the course of one year, from March '25 to March '26. Anthropic banned these accounts from using Claude for violating our Usage Policy. The accounts in this analysis are just a subset of those we investigated and banned during this time period. We selected them because we had enough detail about their malicious activities to map their techniques onto the MITRE ATT&CK framework.

"The 832 accounts in our analysis used AI models for all 14 tactics and 482 unique sub-techniques across the framework, from initial reconnaissance through final impact. We also developed a risk-scoring framework to assess how much AI assistance helped these actors plan their attacks. Most strikingly, we found that the percentage of actors labeled as being medium or high risk jumped from 33% to 56% between the first and second halves of the year." Okay. Let me make sure everybody gets that because this is an important issue. We found that the percentage of actors that they were labeling as being medium or high went from a third, 33%, to more than half, 56%, between the first half of their analysis period and the second half of their analysis period.

They wrote: "This suggests that AI is helping attackers conduct increasingly sophisticated cyber operations with greater ease. Our analysis resulted in three key findings. First: The

number of actors using AI for cyber operations is growing, and their actions carry higher risk. As mentioned above, the percentage of medium- or high-risk actors increased by a factor of about 1.7 in under a year, from 33% during the first half of our study window to 56% during the second. That growth is concentrated in actors using AI for some of the most harmful activities - including lateral movement, credential dumping, and web shells - that carry the highest per-actor risk weight in our scoring, rather than the commodity build-and-obfuscate work that dominates the rest of the population.

"Traditionally, only the most technically sophisticated actors could operate across the entire kill chain, or the sequential stages of a cyberattack. But our analysis found that this is no longer the case. The platform through which they access the model, such as an API or an agentic coding platform like Claude Code, also has no bearing on how high-risk their actions are. What does distinguish the highest risk actors is which techniques they're asking the model to perform."

Okay. "Second: Agentic scaffolding will make it possible for cyberattacks to be far more autonomous. As AI-enabled cyber techniques become more common among this population, it will become harder to differentiate an actor's risk level based on what they're asking a model to do. Instead, the differentiator will become the scaffolding - the surrounding code, architecture, and tooling that makes AI models more capable - that actors build around the model so they can chain together attack stages autonomously. This was starkly apparent in the cyberespionage campaign we disrupted in November 2025, which had a maximum risk score of 100, yet only used a number of techniques comparable to medium-risk actors. That attack was distinct, not because of the number of techniques it employed, but because of how the attackers used an AI agent to orchestrate them.

"Third: The MITRE ATT&CK framework does not yet cover the autonomous actions that make these actors so dangerous. Autonomous kill chain orchestration, real-time pivot decisions, and AI-directed execution with no human intervention do not yet have ID numbers in the ATT&CK framework. Our report included 13,873 observations of malicious activity, all of which mapped to categories laid out in the framework. But the behaviors that distinguish the highest risk actors, and determine the speed and scale of their operations, do not yet have such IDs. The taxonomy that modern threat intelligence relies on must be evolved to capture them.

"While Claude Mythos Preview demonstrates where frontier AI cyber capabilities are heading - models able to find and exploit vulnerabilities at a level approaching the most skilled human researchers - this report tells us how threat actors are already misusing generally available models today. It also serves as a guide to how threat actors are likely to misuse increasingly capable models in the near future, giving defenders a chance to get ahead of them." I hope.

And they finish: "What we learned from this and other analyses directly shapes how we build Claude to prevent such misuse. For example, we've updated the classifiers built into Claude to detect the highest risk actors, and have expanded our probe detections to cover high-risk behavior indicators revealed by this analysis. These findings point to a landscape where the dividing line between low- and high-risk actors is no longer technical skill, but orchestration, and where defenses, detections, and the shared frameworks we all rely on will need to evolve as fast as the attacks they describe."

Okay. So there's so much here. No one who's been following this podcast has ever heard me run around saying that the sky is falling. But what we learn from this report is as close to that as we've ever seen. This extremely sobering report shows that while we've been focused upon and enraptured by all of the many productivity benefits the use of LLM AI can bring to our lives, malicious actors have been exploring the many ways that

same power can be used to attack our world. And unfortunately there are many. The extreme leveraging power of AI cuts both ways.

During the many years of this podcast before AI, our longtime listeners will have often heard me suggest that many major cyber powers must have been assembling, maintaining, and growing a large database of known vulnerabilities. Because we know most of the world is not updating their systems. That database will be large, unfortunately, because there are so many vulnerabilities which we've encountered over the last 20 years. My thought was always that, when a nation-state actor wanted to attack someone specific, they would determine which equipment and versions were being used, then look up the known vulnerabilities in their carefully curated master vulnerability database and launch their attack.

As it turns out, that's not the way it's going to happen. Instead, all of the well-meaning security researchers and software publishers around the world have been publishing, as we have for the last 20 years of this podcast, all of this information for decades. And thanks to AI model training, it doesn't need to be curated into any master reference database. Instead, any well-trained malicious AI will have absorbed all of that knowledge and will have it at the tip of its virtual fingers when it's asked to target a specific entity.

The most important point to appreciate is that bad guys are only using publicly available cloud-based AI, such as Claude, GPT, or Gemini, whatever, because we're still in the earliest days of where this is all headed. I cannot say that enough. I mean, just the fact, Leo, that we see, you know, a new model comes out an hour ago, and it's a dramatic improvement over what we had that came out three weeks ago. I mean, this is just moving so fast. And I guarantee you that, like, it's not like we're running out of steam here. We're still accelerating. You know, AI's legs are just being stretched at this point.

So widely available public cloud-based AI services are currently investing a tremendous amount of time, effort, and resources into erecting, maintaining, and refining the guardrails around their AI because they have no choice; right? They must do everything they can to prevent the abuse of their publicly available services. And the essential nature of LLM-based AI means that even that is not easy.

As we saw, Anthropic is saying, well, you just can't talk to us on our most advanced model about cybersecurity or biotechnology because we're just going to say no. We don't believe that we can determine, you know, well-meaning cybersecurity questions from malicious ones. And in fact there may not be really any difference because, if a security researcher wants to know about something bad, that's the same as a bad guy wanting to know about something bad. They're just going to use the information for different purposes. So you just can't give out the information.

Here's the problem. We already know that AI is able to run quite well off the cloud, locally, on local hardware. It may not be super strong, not like what you have in the cloud with massive data centers and crazy H200 chips that cost, what is it, $40,000, some insane amount of money that NVIDIA's getting for these chips. That's going to change. Locally run AI will have no safeguards, no guardrails of any kind to limit its actions. In the very near future, it will be local models that the malefactors will be employing to direct their real-time attack campaigns. They're not going to be using Anthropic's Claude where it's booting 838 of them, or 832 of them, off of their accounts.

They're going to invest in local hardware just like bitcoin miners did back in the day. You know, the strongest hardware available. And that's going to be running future attacks. And it'll have no safeguards. That's what's going to happen. And while it's against my nature to warn that the sky is going to fall...

Leo: The sky is falling, for sure.

Steve: In the near future I'm not sure I'd want to be spending too much time out in the open.

Leo: Okay, Chicken Little.

Steve: We have far too much legacy mess to clean up and not nearly enough time or incentive to do it. We just don't see people updating the systems that are not giving them any trouble. Even that they are old and bug-ridden and probably already harboring some malware. Well, they're going to get some company.

Leo: Yeah.

Steve: We're in trouble. The bad news is that once attackers move to the use of their own local AI, our ability to monitor their actions, as Anthropic did, at the AI prompt level, it's going to disappear. So we're not going to know what they're using their AI for. We're only going to see astonishingly sophisticated attacks, as if world-class hackers who knew everything about everything were entering systems and pivoting like masters and moving through networks and taking them over. Like I said, it's going to be good times. The good news is that, as Anthropic's year-long study shows, this has not happened yet. So at the moment we're able to see what these miscreants have been up to.

Anthropic writes: "The findings in this report are drawn from 832 accounts that Anthropic banned for violating cyber-related parts of our Usage Policy between March '25 and '26. We identified these accounts through a combination of automated safeguards and investigations by our Threat Intelligence team. For each account, we produced a summary of the observed activity. We then extracted the tactics, techniques, and procedures (the TTPs) described in those summaries, and mapped them to the version of the MITRE ATT&CK framework that was live at the time, which was v18.

"In all, we observed 13,873 actions across 482 unique techniques and all 14 tactics. We gave each actor a risk score from 0 to 100 based on a new methodology we've developed called the AI Risk Enablement Score. So it's ARiES," which they're calling Aries. "We've anonymized the data so that actors cannot be identified in the analysis that follows."

Okay. So I'm going to skip past the description of their scoring system because the most interesting part of Anthropic's report is what they learned of the way threat actors are using, I would say abusing, AI today. And Leo, let's take our last break, and then we're going to look into this breakdown.

Leo: All right.

Steve: What is actually being done by the bad guys.

Leo: And while you've been talking I've been getting all sorts of security work done with Fable.

Steve: Wow.

Leo: Patching all my holes. It's amazing what it's finding.

Steve: And these are holes that previous AI did not know about.

Leo: Claude didn't see - 4.6, 4.7, 4.8.

Steve: And actually, previous AI wrote that code.

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

Steve: So these holes which are security vulnerabilities were created by previous generation AI.

Leo: Good point, yeah. That's a good point, yeah. It's been fun just running through everything with Fable. Fable's very smart and very fast. Very impressive. It's really interesting to see this at work. I think another big jump in capability.

Steve: Just happened.

Leo: It's hard to believe.

Steve: I know. I know.

Leo: It's happening pretty fast here.

Steve: There was an article I read this morning saying that math, generically, mathematicians, it's falling to AI, in the same way that chess did to computation.

Leo: Yeah.

Steve: Basically there are high-end math, theoretical math that has been eluding mathematicians, and AI is now resolving those.

Leo: Resolving those.

Steve: In fact, now it's not producing the proofs. It's now the mathematicians are trying to understand the proof that the AI provided. So the AI says, yeah, here you go. And now the humans are like, what the heck?

Leo: It doesn't seem like this is just pulling stuff out of its knowledge base and applying it. It seems like it's creating new stuff. It's kind of amazing what's happening. I don't know. Anyway, I love following these stories, I do, and will continue to do that more with Steve and Security Now! in just a little bit. Actually we'll learn how the bad guys are using it next.

Steve: Okay. So what did they find? They wrote: "Our empirical analysis of 13,873 observed techniques reveals clear patterns in how adversaries are using AI across the attack lifecycle, and the most common techniques that models are being used for today. The most common technique family we observed was 'Develop Capabilities'" - that's one of the MITRE ATT&CK categories - "'Develop Capabilities' used by 574 out of the total 832 actors in our analysis, which is 69%. The majority of this behavior manifests as 'Malware Development,' used by 560 out of those 574. In practice, we observe threat actors misusing models to build and refine custom scripts to run, write DLL injection code with detailed guidance on how to implement it, as well as canvas fingerprinting evasion and automated account management.

"The next most prevalent techniques are 'Obfuscated Files or Information' employed by 64.7% of threat actors; 'Data from Local System' employed by 55.9%; and 'Impair Defenses' employed by 54.9% also. Together, these top techniques show that threat actors most commonly seek LLMs' help to build pre-engagement offensive tooling, making those tools harder to detect, and harvest data from compromised systems.

"On the other hand, actors are much less likely to use LLMs for real-time, adaptive decision-making" - which that's where the real danger is; right? - "once they've gotten inside a target network." So less likely. "For example, only 54 of 832 threat actors (6.5%) used models for lateral movement, and less than 12 actors used models for remote services like RDP, SSH, and SMB. Only 22.5% of actors used LLMs for privilege escalation and impact stages." So those are all, you know, post-infiltration actions.

"Some technique families that are staples of real-world cyberattacks - such as active directory exploitation, Kerberos ticket attacks, cloud infrastructure manipulation (AWS, Azure, GCP), and container escape, they noted - "have lower representation within the dataset."

So basically the bad guys are sort of still using AI in the old model, not for dynamic real-time attack, but to help them build malware. And that's going to change. One observation I have is that Anthropic's view of what the bad guys are doing is probably skewed by the fact that their AI is deeply wrapped in guardrails. Right? So the lack of more sophisticated use of AI's potential is also probably a combination of the resistance the cloud-based services already have built-in against abuse, coupled with how early we are still in this game.

As I said, it's going to change. And, boy, when they start using local unrestricted models, we're really going to see a change. The strongest supporting evidence for this is the fact that Anthropic noted the large jump in their own risk assessment of what bad actors were doing. Remember it jumped from a third of them to more than half of them, up to 56% of them in the second half, which suggests that we are nowhere near reaching any sort of steady state. This is all still very much growing.

So they continue writing: "The top techniques and the frequency with which actors used them did not change much over the one-year period we studied. For both the first and second halves of the period, the median number of techniques the model is used for is 16." Meaning, right, 16, there are just as many threat actors who used fewer than 16 as who used more than 16. So median as opposed to average.

"In the second half of the year, we observe a subtle directional shift, with threat actors using models less to build standalone malware or obfuscation scripts and more to help with specific operational phases in a cyberattack, and for on-target discovery and collection techniques." In other words, that's where the sophistication is, and that's where the bad guys are going. "Specifically, we observe an 8.9% increase in Account Discovery occurrences, as well as a 6.2% increase in Automated Exfiltration, alongside a 12% decrease in Develop Capabilities, which was that frontend stuff, and a 8.6% decrease in Phishing." Again, more in the post-infiltration and less on the getting ready to do so.

They said: "Defense evasion is the single largest tactic category in the dataset, present in the behavior of 84.4% of the actors we studied. MITRE defines 64 techniques under the broad category of 'defense evasion' across its Enterprise- and Mobile-specific frameworks; we observe 32" - so exactly half - "of these techniques in our dataset: 25 for enterprise and seven for mobile.

"The top techniques observed within this tactic include Obfuscated Files or Information, where 64.7% of threat actors in their sample used AI to implement techniques like XOR or Base64 encoding. So this is obfuscation of information; right? Polymorphic variants, and anti-detection wrappers to evade signature-based detection.

"54.8% used Impair Defenses, where the AI was used to bypass, disable, or tamper with the endpoint security tools, getting around whatever was there to try to catch them. And 30.3% of actors used AI for Process Injection to write malicious code that could be injected into legitimate processes, such as hollowing out processes and DLL injection, to execute payloads from trusted process memory.

"Less frequently used tactics include impact, exfiltration, privilege escalation, and lateral movement. Together, these account for just 8.7% of all observations, less than defense evasion alone." So they said: "Overall, the actors with the highest risk scores used AI most heavily for post-compromise, hands-on-keyboard techniques, such as remote services, credential dumping, web shell deployment, and internal network and account discovery. Lateral movement was the strongest marker of a high-risk actor: 54 actors in their dataset who used lateral movement had an average risk score of 56.4, which was 10 points higher above the average, which was 46.8. No other technique came close to having such predictive power." So lateral movement is what the heavyweight hitters are using.

So the aspect of this entire study that's most unclear to me, again, and I mentioned this before, Leo, is how these researchers avoided the problem of altering the behavior of their abusers. You know, they're describing wide-ranging malicious activity that they apparently directly observed. So does that mean that they allowed Claude to perform these services for the bad guys? Did they drop their guardrails in order to see what the bad guys would do? That's difficult to imagine. But if they did not, then those attempted malicious actions would have been detected and blocked, I would think.

Leo: You would think.

Steve: Yeah.

Leo: You know, in the mine cart for the new Fable, it says we will - we have put in all sorts of hidden things to prevent this. And it will, you know, and they said it's not jail-breakable, which I find to be a very cocky thing to say, and very unlikely.

Steve: Uh-huh.

Leo: But they are definitely trying to keep bad guys from using these models for malicious code. Whether they'll succeed is another matter.

Steve: So they said: "What this means for defenders," right, so here we are, the good guys. What does what they saw mean? They wrote: "The population of AI-enabled actors is not only growing, but also drifting toward the riskiest activities in our framework, without requiring the actors themselves to become any more skilled." So there's the gotcha; right?

Leo: Yeah.

Steve: That's the danger is that it has lowered the bar of skill level, so less skilled, and of course it's a pyramid; right? There are many more less skilled threat actors than there are those, the cream of the crop at the top. They said: "If this trend continues, these operational techniques will not be a differentiating factor anymore" - that is, you won't be able to tell the skill of an actor because they'll all be doing the fancy stuff - "and will become the baseline for tomorrow. And we'll need to find a new way to measure the riskiest actors." I don't care about measuring them. I mean, they're bad.

Okay. They said: "Looking at our highest risk threat actors also underscores that calculating the risk of AI-enabled cyber operations based on number, type, or breadth of attack techniques is insufficient." Yeah, or I would argue irrelevant. "We need a way to understand the scaffolding threat actors are able to build to chain these techniques together." And Leo, I just see this being automated. I see it being sold on the dark web. There will be scaffolds which the advanced guys sell to the junior guys...

Leo: Exactly.

Steve: Which automate all of this for them.

Leo: Yeah, yeah.

Steve: And they said: "This will allow them to use AI models to autonomously execute large swaths of a cyberattack without human intervention." Now, get a load of this. Here's the one guy. They said: "We analyzed the behavior of the threat actor who orchestrated the AI-enabled cyberespionage campaign we reported on in November 2025." They labeled this threat actor GTG-1002. "We see that this actor achieved the maximum possible risk score of 100" - remember that the average was down at 46.2, so this guy, this was an elite, you know, group - "successfully compromised government and critical infrastructure targets across multiple countries, and developed a scaffolding to use Claude Code, not as an advisor, but as an autonomous operator.

"Yet their overall MITRE profile - 30 techniques across 13 tactics - is comparable to dozens of medium-risk actors in this dataset. The median actor deploys 16 techniques." So they were even below the median. "Several low-risk actors also exceed 30. In other words, technique count or tactic type alone could not explain what made GTG-1002 the highest risk actor we've observed thus far.

"What does explain this actor's high-risk score is the increasingly agentic components they used: how they were able to orchestrate and chain together techniques to take action on their objectives. GTG-1002 weaponized Claude Code running on a Kali Linux machine, integrating open-source penetration testing tools as MCP (Model Context Protocol) servers, effectively turning the AI into an autonomous attack platform rather than a code-writing assistant. The AI didn't just suggest commands or generate attack scripts, it executed them and reasoned about attack environments autonomously. Some indications of their 'agentic-ness' show up proxied through the types of techniques we track. GTG-1002 employed operational techniques such as 'Remote Services: SSH,' 'Exploitation of Remote Services,' and 'Archive Collected Data.'" Those are the MITRE categories.

Their analysis concludes with a very clear description of what they observed this most advanced threat actor, which they codenamed GTG-1002, doing. And if you want a chill to run up your spine, just remember that we have barely begun, and that what may be a single top-rated risk profile actor today will almost certainly become every threat actor once the understanding of how to best leverage these tools becomes widespread. And we know it will become widespread.

So Anthropic explains what this one actor, GTG-1002, did. I have three bullet points. "First, autonomous execution within stages: GTG-1002 deployed Claude Code running on a Kali machine to orchestrate dozens of MCP tool operations autonomously - scanning and mapping dozens of Internet-facing services during reconnaissance, then discovering internal admin portals, databases, logging servers, and temporal workflow systems once inside the network. The AI didn't just suggest commands; it executed them, making tactical decisions about what" - I'm getting goose bumps - "making tactical decisions about what to probe next without waiting for operator input.

"Next, live exploitation and pivoting: Operating within GTG-1002's scaffolding, the AI exploited an SSRF (Server Side Request Forgery) vulnerability in a public-facing web server to proxy commands into the internal cloud environment, harvested SSH private keys from internal infrastructure and service account tokens from cloud metadata services and AWS Secrets Manager, and used those harvested credentials to move laterally across the victim's cloud environment. These are the operational phases (discovery, credential access, lateral movement) that were more rare in our dataset.

"And finally, human intent, AI execution: GTG-1002 provided strategic direction while the AI handled tactical implementation. The AI operated autonomously" - I can't believe I'm even reading this, Leo. A year ago this was sci-fi. I mean, it was. A year ago this was science fiction.

Leo: Yup.

Steve: And it is now real. "The AI operated autonomously during reconnaissance and internal discovery, adapted its approach when it encountered unanticipated infrastructure like container image signing workflows and service account identities. It staged and compressed tens of thousands of proprietary workflow records and internal architecture documentation for exfiltration. The final data extraction - downloading to the attacker's machine via cURL MCP tool calls - was human-directed, suggesting the operator retained control over the consequential decisions while delegating the operational work to the AI.

"GTG-1002's activity was novel for using an AI agent to autonomously chain together many stages of the cyberattack lifecycle - reconnaissance, exploitation, lateral movement, and exfiltration - into a coherent operation, making real-time decisions about what to do and what data to collect. This is the dimension of AI-enabled uplift that a

technique-frequency table cannot capture, and it is the dimension we expect to matter most as agentic tooling matures."

So while the good guys are currently excited about the promise of using the awesome leverage of agentic AI to transform our lives for the better, we should also soberly recognize that malicious forces throughout the cyberworld are every bit as excited by the capability they are now receiving to dramatically magnify the power of their cyber attacks. This new AI technology with agentic agents coupled with MCP to remotely control existing tools is, you know, it's neither good nor bad in itself. What it is, unfortunately, is a great deal of both.

Leo: Yeah. And we've got it. Nothing you could do about it now. It's here.

Steve: No, but, I mean, but please, take...

Leo: It walks among us.

Steve: Everybody take this seriously.

Leo: Well, I think today's the day because of the release of Fable. I think it is now. It's here.

Steve: Right.

Leo: It's not as fully capable as Mythos, I guess, but I'm really impressed by what I've seen so far. It's pretty amazing. Steve Gibson. Go ahead.

Steve: I was going to say, so Fable is not Mythos? It is a...

Leo: It's Mythos-adjacent. [Crosstalk] Claude 6. It's the next Claude model. But I think it's very related to Mythos, yeah. It has certainly security capabilities. But Mythos, remember, wasn't trained specifically for security capabilities.

Steve: No.

Leo: It's just a really good model.

Steve: Right.

Leo: That's what's happening, yeah. Fable. Claude 5. Not Claude 6. I skipped one. The last Claude 4.6. 4.5, 4.6, 4.7, 4.8, and we're now at [crosstalk] 5.5.

Steve: Well, and they're calling it Fable 5.

Leo: Yeah.

Steve: When it popped up, it was Fable 5.

Leo: 5, yeah.

Steve: Yeah.

Leo: Exactly. I'm sure, I mean, this just - it just came out. This is - we're going to see a lot more. But so far the buzz is very, very positive everywhere I look.

Steve: Wow.

Leo: Which it wasn't for 4.8. 4.8 was pretty much universally reviled. But I think 4.8 was an interim release because they wanted something to fall back to when they released Mythos.

Steve: And, you know, when you said this, I was reminded of how Windows versions always alternate between good and bad.

Leo: Oh, yeah.

Steve: You know? It's like...

Leo: Maybe. Maybe, yeah.

Steve: Yes, I don't know why, but like every other Windows major release was...

Leo: That's right, it's famous for that, yeah.

Steve: Yup.

Leo: Famous for that.

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/