HalluSquatting, GhostApproval & GitLost

Description: Europe warns their largest banks to prepare for AI attack. The EU launches an action plan for AI Cybersecurity. China considers keeping its budget AI to itself. The UK's NCSC and GCHQ announce their "Cyber Shield." CISA is using Mythos to audit U.S. government code. Microsoft warns of their upcoming patch flood. "RoguePlanet" receives an on-the-fly patch. An underused mode to kid-proof an iPhone. Listener feedback and three new AI attacks.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1087.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1087-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. Massive Patch Tuesday for Microsoft. A thousand-plus fixes. We'll talk about that. We'll also talk about some interesting hacks against AI that you might be aware of or want to be aware of. And Steve has found a really good way to kid-proof an iPhone, or make it suitable for adults who just don't need all those icons. That's coming up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1087, recorded Tuesday, July 14th, 2026: "HalluSquatting, GhostApproval, and GitLost."

It's time for Security Now!, the show where we cover the latest in security, privacy, online stuff, computers, vitamins, and this guy right here, Mr. Steve "Tiberius" Gibson.

Steve Gibson: And these days, my friend, the world as AI.

Leo: It's fascinating.

Steve: Not that there's, I mean, I'm now over feeling sheepish about basically just talking about AI on this podcast because AI has impacted cybersecurity like nothing else ever has. Truly. We're going to talk about several nation-states' declared initiatives about the impact of AI and their own cybersecurity - the EU, the UK and China. The UK has got something. It's interesting to me that in some ways they're leading us, because I've seen nothing like this from us. But the UK's NCSC and GCHQ, the National Cyber Security Center and their main intelligence group, they've announced something called Cyber Shield, where they're talking about, I mean, this is Colossus. They're talking about turning their national cybersecurity over to autonomous agents.

Leo: Oh, that's not - that's scary.

Steve: And laying out the whole protocol. Anyway, we're going to talk about...

Leo: When you say "Colossus," you're talking about "Colossus: The Forbin Project."

Steve: The Forbin Project.

Leo: For those who don't know.

Steve: Yes, yes.

Leo: And it is on YouTube. You can watch it on YouTube if you have...

Steve: I've not looked at Patch Tuesday, but this is Patch Tuesday, second Tuesday of the month. And Microsoft has warned the entire industry in a blog post that we'll look at to expect many more patches. Of course, we anticipated this months ago. But now they're saying, oh, we counted them. Whoops. Anyway, so, good. We'll talk about that. RoguePlanet, the latest salvo from Nightmare Eclipse, has been fixed on-the-fly with a patch. We'll talk about that.

Also just a random tip that I ran across in WIRED about a much better solution for kid-proofing an iPhone that I just had to share with our listeners because I've heard so much feedback from our audience saying, you know, like questions about how do they childproof their Internet connection? What do they do in order to keep their kids from going places they shouldn't? So I know that that's a strong interest in controlling what youngsters do. And this was just a really cool tip from WIRED.

Then we're going to talk about three new means of subverting AI. Something known as HalluSquatting, as in hallucination squatting. This is a really cool hack. We've got a second one called GhostApproval. And then my favorite name for a hack in a long time is GitLost.

Leo: GitLost.

Steve: GitLost.

Leo: Very nice.

Steve: So anyway, the title for today's Security Now! 1087 for this 14th day of July is "HalluSquatting, GhostApproval, and GitLost."

Leo: I love it. That's all coming up, all three and more, in just a little bit.

Steve: We do have another puzzling Picture of the Week also, Leo.

Leo: I see, you know, it's funny, normally I don't, you know, I can't, I don't look at them ahead of time because I like to be surprised along with everybody else. Well, everybody doesn't get the email. But I can see the top of this picture, and I have a feeling I kind of know what the bottom is going to be. But we'll see. We'll get there. I am ready for the Picture of the Week. Oh, you're muted. What? Hello? Did I mute you? You're muted. There we go.

Steve: Sorry. So this picture was taken by a listener of ours, and we're getting more of those because people sort of have the podcast in mind now, and then they see something wacky, and they go, okay, we have to share this with Steve because maybe this would be worthy of the podcast. This is in Australia. Our listener works for IBM, and this is an IBM facility which he said IBM owns. I gave this the caption "Some building safety officer may have required the sign, but how could this have happened?"

Leo: I did see the top, and I thought, okay, that looks like a staircase to nowhere. And the sign says "No Access Beyond This Point." Obviously.

Steve: One of our clever listeners responded. The email to about 20, almost 21,000 of our listeners went out late morning on Sunday. Someone wrote back and said, "Well, I'm not a Muggle." So, you know...

Leo: Is this Platform 9-3/4? Are we going to walk right through that ceiling?

Steve: Or I'm sorry, he is a Muggle. He said, "I am a Muggle." Meaning that he isn't able to walk through. But Leo, I - obviously there was once no ceiling here. So, okay, I'm sorry. I forgot to tell people who are not looking at this why we're kind of gobsmacked, as they say in the UK, I guess. We have a staircase which is winding around as it goes up floor by floor. You know, like a floor of stairs, then you turn a corner twice, come back to where you began on the floor above, and then go up again and so forth. Which goes into the roof, into the ceiling of where this is.

So to me, one would expect that, okay, once upon a time there was no ceiling there; right? Like some renovation at - I just can't explain it. Oh, okay. So if there was a staircase there that continued to go up, then that would have bisected those two floors with the stairs. So if they were to join the floors that the stairs were going up to, then I guess you'd have to put a floor across where - anyway, I can't...

Leo: So there are some theories in our chat.

Steve: Okay, let me hear them.

Leo: And one would be, now, this is an IBM facility. One would be that they wanted to limit access to this next floor. Maybe it's a secure floor. It costs more to take the stairs out. It's very easy just to seal up the entry. So they just put the floor in, and they didn't bother taking the stairs out. You can see there's stairs below it that led up to this floor.

Steve: Right, right.

Leo: So this is a staircase that goes up. The other theory, Darren Oakey, who lives in Australia, says he's seen this several times, is that a company rents several floors in a building, but maybe they didn't need that top one. So they just...

Steve: Ohhh.

Leo: ...closed it off so that somebody else could rent it. But again, taking staircases out, especially if it's a staircase that, if it were just on this floor maybe that'd be easy to do. But you can see it's part of a staircase that goes below.

Steve: Yes, yes. So the idea would be that they said, okay, we're going to give back this floor above.

Leo: Right.

Steve: Or, like, give it to someone else.

Leo: Right.

Steve: So they literally just, you know, terminated their access.

Leo: Yeah, they just put a big thing across it. You don't want to go there.

Steve: And of course then, Leo...

Leo: But why do you need the sign because it's so obvious?

Steve: The sign. I forgot to tell everybody who isn't seeing it. The sign says no access beyond this point.

Leo: Unh-unh.

Steve: And then with a little IBM, you know, their standard logo.

Leo: You know what, maybe you've got a distracted employee, he's not paying attention, he's hustling upstairs.

Steve: He bumps his head.

Leo: He doesn't look up, and he runs right into the - so maybe there's a...

Steve: And as I said, and thus the reason for the caption I gave it: "Some building safety officer may have required the sign."

Leo: Exactly. I love it. That's hysterical.

Steve: Okay. So last Tuesday TheNextWeb picked up the story of the ECB's (European Central Bank) recent AI warning to Europe's entire banking industry. And again, I'm sure our listeners by now know that AI is impacting cybersecurity probably, I would argue, more than anything else. Can you think of, I mean, we know it's impacting coding, yes. But, I mean, and there's sort of an expectation of it having an effect throughout other industries; right? Like the law profession getting rid of, like, low-end sort of middle manager attorneys whose jobs can be now done by AI. But to my way of thinking, though I actually have some tunnel vision on this, I can't think of anything where AI has had such an immediate, like, right now, not next year, not coming soon, but, I mean, right now effect, as in cybersecurity. To me it seems like it's the thing that AI has impacted most immediately.

Anyway, TheNextWeb picked up the story about this, saying that: "The European Central Bank has a warning for the euro area's largest banks. Frontier AI now poses a serious cyberthreat. And again, not in the future, today, like immediately, like the moment it happened. The lenders (banks) must draw up plans to counter it. Claudia Buch, the chair of the ECB's supervisory board, wrote to bank chief executives setting a deadline for the end of October. Buch asked lenders to patch software faster and harden their AI-enabled cyber defenses." The idea they even have them already is something.

"She also wants tighter oversight of the outside technology partners they lean on. Over the longer term, banks must modernize ageing infrastructure and sharpen their crisis response. The order carries no fines. Banks that ignore it face no formal sanction." So literally just a warning. Just a wakeup call, but with no teeth. "The ECB," they wrote, "says it may still use the plans to rank lenders against each other and press the laggards.

"One model looms over the letter." And I'll be talking about this in a second. "One model looms over the letter: Anthropic's Claude Mythos. Anthropic says Mythos can find unknown flaws in IT systems. It claims the model has already spotted thousands of severe vulnerabilities across major operating systems and browsers. The company first limited who could use it, which spread unease across European finance.

"Buch put the worry plainly, she wrote: 'Emerging models can pinpoint software weaknesses and write working exploits at unprecedented speed. That collapses the gap between finding a flaw and firing through it.'" That's an expression I hadn't heard before, "firing through it," but okay.

"The ECB did not act alone. The same day, the European Systemic Risk Board lifted its assessment of systemic cyber risk to 'severe.'" So they've got some sort of a DEFCON scale. And so we're now at severe systemic cyber risk, saying that frontier AI should now count as a source of systemic risk in its own right. "The board also flagged a second worry. Nearly all the leading AI providers sit outside the European Union. That leaves the bloc dependent on foreign firms and exposed to geopolitical pressure. It urged Europe to build its own AI muscle." Which we're going to see this here in a couple seconds with some of the follow-on stories. AI is now, and should be, and so not surprisingly, being

viewed as a strategic resource, not unlike the power grid. I mean, you have to have it. And you have to know, you have to have a reliable supply of it.

"The move," they write, "fits a wider European scramble. ECB President Christine Lagarde warned last month that AI could trigger a dangerous financial crisis. The ECB has already run 109 banks through a severe cyber-attack drill. The threat is not hypothetical either, with state-backed attacks climbing and European bodies such as France's statistics office already hit.

"A market has sprung up around the fear. French startup Mistral has opened talks with European banks to sell a flaw-hunting tool." So now we've got, you know, somebody local in France who is beginning to say, okay, we've got some commercial capabilities. They wrote: "It's one of several firms racing to offer a homegrown answer to Mythos. For now, the regulators sound clear on the danger, but are much more vague on the fix. Banks have until October to show they're ready."

So this reaction from Europe, of course, as I said, it should not be and is not surprising. But given the rapid uptake in the use of AI for cyber attacks, giving banks until October, to me, you know, what, July, August, September, October, so four months, does seem like an overly slow response. On the other hand, these are banks, so it might be as fast as they're able to move. And the ECB may be painfully aware of that.

The other thing that caught my eye was the specific mention of Mythos. You have to give Anthropic credit for the brilliance of their rollout of this. The skeptics who cried foul that Mythos was just marketing, well, they were wrong. It's not just marketing. But they certainly had a point. Mythos was able to deliver the goods first. It would have been and would still be, still is, truly dangerous in adversarial hands, and Anthropic did take advantage of that for marketing. We've learned since then - seems like a long time ago, but it was like a month - that other less powerful models, when carefully orchestrated, and I heard you say, Leo, over on MacBreak Weekly that it's turning out that the harness that one uses for AI is as important as the model that it's harnessing. And that totally makes sense, and I agree with that.

You know, we've learned that less powerful models, when carefully orchestrated, can deliver similar results. And this was what so miffed those guys over at AISLE. Remember A-I-S-L-E. They had actually preceded Mythos in their own work without the Mythos model behind it, and achieved similar results because they were commercializing this capability. And along comes Anthropic and steals all of the glory. On the other hand, they didn't have the presence or the marketing to pull off the coup that Anthropic managed. And we should remember, in time this will all just be a note in history books. And at the rate we're moving, that'll be a year from now. It'll be like, oh, remember back then? Right.

Okay. So speaking of what's going on across the Atlantic, or the Pacific, rather, last week also the European Union published their so-called "Action Plan." So this is the EU broadly as opposed to just the European Central Bank. ECB was the previous report from TheNextWeb. The EU, their Action Plan, they titled "The Action Plan on Cybersecurity and Artificial Intelligence." In their announcement of this, the European Commission themselves wrote: "Artificial intelligence is rapidly transforming the cybersecurity landscape. AI can help detect vulnerabilities, prevent cyberattacks, and strengthen the protection of critical infrastructure. At the same time, it can also be exploited by malicious actors to automate attacks, identify weaknesses, and carry out cyber operations at unprecedented speed and scale.

"Building on the EU's existing legal framework on AI and cybersecurity, the Action Plan" - that's capital A, capital P - "the Action Plan sets out a coordinated approach to help Member States, businesses, and public authorities benefit from the opportunities offered

by AI while addressing the new risks it creates. It focuses on three complementary objectives: Promoting the safe and responsible use of advanced AI; reinforcing the EU's cybersecurity and resilience; and scaling up Europe's AI capabilities for cybersecurity.

"To promote the safe use of advanced AI, the Commission will strengthen Europe's capacity to evaluate AI models before they are placed on the EU market." Essentially they're saying, what? This caught all of us by surprise. We need to figure out how to even know what an AI model is, what it does, is it good, which one should we choose. So they're saying we need to "strengthen that capacity to evaluate AI models before they're placed on the EU market, in line with the AI Act." They said: "It will also work with the European Union Agency for Cybersecurity (ENISA) to develop a European Blueprint for secure access to advanced AI systems for cybersecurity purposes and establish a secure testing platform to help organizations in critical sectors, such as energy, transport, health, finance, and public administration, safely test and deploy AI solutions."

So here again, the message is we've been sort of focused on companies like Anthropic, well, and their interaction with the U.S. government specifically, in sort of a microcosm. But imagine that you're a nation-state, and this is happening outside your borders, yet is clearly going to affect entities, your existence, your network, your national security. So suddenly you're expected to be able to deal with this bureaucratically the way you deal with everything else bureaucratically. I mean, this is just a whole new thing that requires a massive new bureaucracy in order to, like, figure out what to do about it.

They said: "The Action Plan also reinforces the EU's cybersecurity by promoting the implementation of existing EU cybersecurity legislation, including the NIS2 Directive and the Cyber Resilience Act. It encourages organizations to use AI, including open-source models where appropriate, to detect and address vulnerabilities more quickly and improve their ability to prevent and respond to cyberattacks." In other words, everybody's got to get going here in the EU.

They said: "To strengthen Europe's technological leadership, the Commission will launch an EU Grand Challenge on AI for cybersecurity, bringing together companies, researchers, and other stakeholders to develop innovative AI-powered cybersecurity solutions. The EU will also continue investing in sovereign AI capabilities" - meaning, you know, their own, we need our own AI they're saying. China's got it, the U.S. has got it, what happened? We don't got it. So "sovereign AI capabilities, building on initiatives such as AI Factories [whatever that is] and future Gigafactories, while encouraging private investment to help scale up European AI technologies." In other words, we got kind of caught flatfooted here. We need AI.

"The Action Plan," they finish, "complements the EU's existing legal framework for AI and cybersecurity, including the AI Act, the Cyber Resilience Act, the NIS2 Directive, the Digital Operational Resilience Act (DORA), and the Cyber Solidarity" - they've got a lot of acts over there.

Leo: Yeah. Might be part of the problem.

Steve: You know, not much is happening, but boy have we got some initiatives, Leo. We've got acts.

Leo: We'll make a committee to talk about this one.

Steve: That's right. We're going to, well, you have to have a committee to talk about the committee.

Leo: Yeah.

Steve: You know.

Leo: Who's on the committee?

Steve: What should the shape of the committee to do this take? And that's where they are on the whole child age dating thing, too. The W3C had to - they're at the stage of fashioning the committee that will then begin working toward - it's like, just do it, guys. This is not hard.

Leo: Drive me nuts. Because they think they just - all you have to do is write a bill or a law or an executive order, and we'll create our own local AI. But, you know, there are companies in Europe that've been trying - Mistral is in France. They're not competitive.

Steve: No.

Leo: You can't just make it with the stroke of a pen.

Steve: No.

Leo: It's just bizarre.

Steve: And look at the investment. I mean, now, of course, in the U.S. we have the data center controversy because it's upsets so many - there's a piece...

Leo: New York just banned data centers, by the way. The whole state.

Steve: There was a piece in the news today that said that someone had been kicked out of their home because, what is it called, I forgot now, the term...

Leo: Eminent domain.

Steve: Eminent domain, thank you, yes. Eminent domain required that a new power distribution system be put through 300 homes which, because in order to deliver enough power for a data center somewhere. So it's like...

Leo: Hey, you know, the Nevada or the Lake Tahoe regional power authority has announced that they're going to cut off domestic users of power. You're going to have to get a new power company because we're going to give all our power to AI. No wonder people are pissed. I understand, I really do.

Steve: I do, too. And then we had the initiative of, don't worry, we're not going to power from the grid. We're going to power locally because there's a huge natural gas pipeline running past us. Well, turns out when they fired up the turbines, people can no longer hear themselves think outside because the whine of the gas turbines.

Leo: Well, yes.

Steve: It's like, oh, well, you know, yes, we have great air conditioning inside, and we have power. But, you know, our children can't play outside.

Leo: This also is because of a short-sighted policy on sustainable energy after years and years and years.

Steve: Yep. Well, how many times have we been talking about how obsolete the U.S. power grid is.

Leo: Right. Right.

Steve: This is what happens when you suddenly need power from the power grid which has been obsoleted.

Leo: There is more than enough power being focused on the Earth from the Sun to do all of this.

Steve: Yep. Yep.

Leo: But we have decided we want to dig up fossil fuels instead.

Steve: And boy, has solar become cheap in the last, I mean, like a while ago the story was, oh, it's too expensive. Well, not, it's really not anymore.

Leo: And this is why China is going to leapfrog us because they don't...

Steve: Right.

Leo: You know, they don't have this problem. Oh, well.

Steve: So what's interesting with all of this, it occurs to me that the initiatives such as the EU is talking about are another less obvious downstream benefit of the way Anthropic marketed Mythos. Just as it shook up the U.S. government, in today's highly connected world it similarly, now, they're looking at it from a distance; right? But it shook up every other nation. The EU's announcement, you know, reads, as it is, like bureaucratic boilerplate. And as you said, Leo, that's what bureaucrats produce is these sorts of acts, this act and this act and the other. So anyway, what it does show us is a view into what those bureaucrats are thinking. And the bureaucrats in the European Union are thinking, "Holy crap, we don't want to get left behind."

Leo: Right. Can't blame them.

Steve: And that brings us to China.

Leo: Oh, boy.

Steve: Reuters' headline reads: "Beijing" - get this - "is looking at curbing overseas access to China's top AI models."

Leo: Uh-oh.

Steve: Uh-huh. I knew that would get your attention.

Leo: Well, people are saying download all the models now before they cut them off.

Steve: Yeah.

Leo: That's the problem is, you know, HuggingFace and others have these models. It's just it would only be future models, I guess.

Steve: Correct. Well, or running the models. So you need to be able to run the models on some powerful enough infrastructure for it to work. And Leo, the other problem is, as you said "future models," I cannot stress enough the degree to which we are hardly started. I mean, every bit of my intuition says that, I mean, just the fact that look what happens in one month. No mature technology has what has happened in a month happen. So this isn't, I mean, it's not even premature technology. It has just begun.

Leo: It's mindboggling.

Steve: It is. It is astonishing. So Reuters reports: "Three people familiar with the discussions said that Chinese authorities have held meetings with top tech firms" - meaning theirs - "over the past month about potentially restricting overseas access to China's most advanced AI models. The talks follow a number of steps by Beijing to keep homegrown AI within the country and underscore how China, like the U.S., is now treating cutting-edge artificial intelligence as a critical national asset that needs control.

Companies present" - I mean, it is. This is what's got the EU all up in a bother is that AI is now a national asset. "Companies present at the talks included tech giants Alibaba and ByteDance, as well as the startup Z.ai, said the people, who were not authorized to speak to media and declined to be identified.

"Since the emergence of DeepSeek's R1 model last year," writes Reuters, "Chinese AI models have made big inroads globally thanks to their low cost and increasing capabilities. Any decision by Beijing to limit access to those products could ripple across AI markets as costs for many businesses would likely increase.

"At the meetings, led by China's Ministry of Commerce, participants discussed putting limits on the most advanced AI models, both closed source and more open versions, according to two of the sources. Officials talked about making any leak or theft of proprietary AI technology an offense under China's stringent national security laws. The officials also raised the possibility of implementing new measures to restrict who can fund domestic AI startups." Meaning probably no foreign funding. "The scope of the potential restrictions is still being discussed and may only apply to future models. It was not immediately clear when or even if they would come into force."

The point being here, this is a discussion point. And this indirectly demonstrates the thinking that is going on. It's like, wait a minute, we, China, have an asset. Notice also, Leo, they don't have a problem creating as many data centers as they want to, in the same way that they don't have a problem flattening land in order to put up solar arrays in order to power them.

Leo: Right, right.

Steve: And at some point, here's the U.S. getting in trouble because it's unable to produce the compute power that suddenly, like in no time, we need. China doesn't have a problem generating as much compute power as they decide that they want to produce. And then that becomes a strategic asset of theirs. Which, I mean, which we have a problem duplicating because we've got too much pushback about the creation of data centers. I mean, thinking of AI as a strategic national asset suddenly puts a whole different complexion on it.

Leo: Yeah, yeah.

Steve: Reuters said: "China's commerce ministry, which oversees export regulations, and the National Development and Reform Commission - the country's state planning agency whose officials also attended the meetings - did not respond to Reuters' direct requests for comment. Alibaba, ByteDance and Z.ai also did not respond to Reuters' queries.

"All three companies have a range of AI models, some closed-source while others are open-weight, meaning users can download, run, and customize the underlying systems. Alibaba's Qwen and ByteDance's Doubao are two of the most widely used AI models in China."

Leo: Qwen's fantastic, yeah.

Steve: "Z.ai has recently set Silicon Valley abuzz as the capabilities of its GLM-5.2 model come close to leading U.S. offerings at a fraction of the cost."

Leo: I use it, and it's about a quarter of the cost, yep.

Steve: "U.S. President Donald Trump's administration has also been deeply concerned about national security implications of AI, in particular the potential for American AI products to be misused by military intelligence in China, Russia, and other countries of concern. In June, it ordered that foreign nationals not have access to Anthropic's most advanced Fable and Mythos models, which prompted the company to disable the models for all users globally as nationality could not be verified in real time. Export controls for Fable, which is designed for the general public, have since been lifted after new safeguards were put in place. But Mythos, designed for cybersecurity professionals, is still only available to some trusted U.S. organizations." And we'll be talking about that in a second.

"Some U.S. AI experts have also said the U.S. needs to regulate the use of Chinese AI models." Yeah, right, we don't like DJI drones, so why are we going to trust some Chinese AI model? "According to two of the sources, Chinese authorities are deeply worried about the potential for Mythos to exploit software vulnerabilities, and that Washington might deploy the model against Chinese interests. That echoes concerns publicly voiced by state media and Zhou Hongyi, founder of cybersecurity firm 360, a major vendor of government and enterprise clients, who has said China needs to develop its own Mythos." Well, we know they are working on that. They have to be.

"This year, China has implemented numerous measures to protect homegrown AI. In April, the country's state planner ordered Meta to unwind its $2 billion acquisition of Chinese-founded AI startup Manus. In early June, authorities issued sweeping new rules, tightening control of overseas deals that involve Chinese investors, technology, data and national security." In other words, walls are being put up. "China had also launched investigations this year into Manus and other local AI startups that had moved abroad, seeking to establish whether they have broken export control laws, according to two of the sources and a third person. Manus has not responded to requests for comment by Reuters.

"Reuters was not able to learn how any potential new restrictions on overseas access to Chinese AI models might work. But some hints might be gleaned from a May roundtable of Chinese legal experts on regulations governing open source AI. According to a summary of the discussions published in an official Supreme People's Court journal, participants proposed a tiered system: basic open-source tools subject to a simple filing, more advanced technologies facing security reviews, and the most sensitive frontier models barred from public release or restricted to domestic use."

So none of that, at this point, should come as any surprise. Right? But I wanted to share this reporting from Reuters, following the news from Europe, to highlight the fact that, as I said at the top, every nation-state is waking up to the fact that what has been happening with AI is not hyped-up marketing and overstatement. I've recently seen the impact of the emergence of AI being compared to the harnessing of electricity. They actually said the invention of electricity; but, you know, it wasn't invented. That doesn't feel like an exaggeration to me, Leo. I mean, it is that big. This is - I don't have words to just explain how big this is going to be.

Leo: Yeah.

Steve: The entire world is being turned upside down by the emergence of this new AI technology. And again, I am utterly certain that we're still at the beginning of this.

Leo: Unless governments and NIMBYS and others kill it. I mean, that's the problem, is...

Steve: Well, they can restrain it. I mean, so if you were to compare this to the nuclear bomb.

Leo: Right.

Steve: The benefit, there's a much greater ability to control the making of a bomb because it requires such physical infrastructure in order to refine the raw materials that go into creating it. And then you've got a delivery system that is also very much in the real world. AI is way more difficult to control.

Leo: Yes and no. It's so capital intensive to make these models. That's part of the reason. It's only these giant companies raising money.

Steve: Today, today, today.

Leo: You think we'll be able to do - but for instance, the estimate right now is that Fable, which is I think the strongest model we have, is a 10 trillion parameter model, one you could not run, obviously, on anything besides a giant data center with almost, you know, infinite amount of GPUs and RAM. It takes - it's capital intensive, kind of in a similar way atomic energy is capital intensive. Or fabs are capital intensive.

Steve: Right.

Leo: You're not going to have garage microprocessor makers. And so that's kind of a gate. Maybe the - what is interesting is there is development along the lines of how do we make, as you pointed out, you love the idea of purpose-built narrow capability AIs.

Steve: Yes.

Leo: Those don't have to be 10 trillion parameters.

Steve: No, no. You can ask Fable anything.

Leo: Right.

Steve: But if you're going to have an AI that specializes in coding, it does not need to be able to talk about the fall of the Roman Empire.

Leo: Right.

Steve: So that knowledge is in the model, and it's taking up space and power and time and compute. And so I am, as you said, I am absolutely certain that we will. And the fact that we don't have it yet demonstrates just how early in we are. There will be application-specific models.

Leo: Well, I'll give you a good example.

Steve: That blow away a general model for way less compute.

Leo: I could use Fable or some giant frontier model. I have all my cameras hooked up to my AI. I could use Fable to analyze the images. I don't need it. I can run a tiny, relatively tiny Qwen model. There's a Qwen visual recognition model. I'm running it with other models in my framework. It's maybe 27GB, I don't know, it's not very big. And it's very good. It analyzes every picture that comes in. It gives me text messages, so there's a person in a yellow shirt with a red hat and shorts, carrying a brown package. I mean, it's that, and it's very, very good. And that's all it does. It's just vision recognition. And we do that on our phones locally, so we know we can get small models to do that.

Steve: Yeah.

Leo: So you're right. If we are willing to slice it up like that, I think you will see garage models.

Steve: And it'll be at a targeted market.

Leo: Oh, yeah.

Steve: A general Fable public model needs to be able to...

Leo: Talk about the Roman Empire.

Steve: ...pontificate on any topic. Right.

Leo: I asked it - what a waste of CPUs. I asked it what the best spray starch would be for ironing my handkerchiefs. And it came up with it. But probably not the best use of all of that horsepower. Anyway, I'm sorry to interrupt. Go ahead.

Steve: No. It's your turn to interrupt. You know what a good use of all this horsepower would be.

Leo: You know what they call this in the World Cup in the soccer matches, this is your hydration break. Steve Gibson. On we go with the show. On with the show we go.

Steve: Okay. Now we get to the UK. The NCSC is the UK's National Cyber Security Center. Their blog posting last week was titled "Cyber Shield: The path to" - get this, Leo - "an agentic AI future for cyber defense. Why the UK is pioneering an initiative to develop a national scale, sovereign defense capability." So again, the implications of AI are, I mean, I guess I think the Mythos drama with the U.S. served as a catalyst to at least make these initiatives public. They were probably in the works for some time. But it's like, okay, now's the time to talk about this.

The posting led by quoting the director of GCHQ Anne Keast-Butler during her recent inaugural GCHQ Annual Lecture, which is held at Bletchley Park. This occurred on May 27th when Anne said: "We need to reimagine cyber security in the AI world. In the past few months, GCHQ has developed a blueprint for a new national cyber defense capability that will hardwire cutting-edge agentic" - and, you know, there's no way to read "agentic" other than autonomous; right? - "agentic AI into machine speed cyber defense."

Okay. So that was quoted at the top of this NCSC posting, which then followed that by writing: "The NCSC and the Department for Science, Innovation and Technology (DSIT) are developing this blueprint, which we are calling Cyber Shield. The objective of Cyber Shield is to build a national-scale, collaborative approach to agentic cyber defense, using frontier AI to identify, reduce, and resolve our national cyber risk.

"In this blog, we set out why we need a new approach to cyber security, and outline our aim and vision for Cyber Shield, as well as the challenges. We also invite debate and engagement from academia, Critical National Infrastructure organizations, frontier labs, and cyber defense sector and others to help collectively solve the challenges and develop the blueprint.

"Why the UK needs a step change in cyber defense: The UK faces a cyber threat that's growing in scale, speed, and sophistication. Attacks from hostile states, organized crime, and others are increasingly disrupting services, harming businesses, and exposing sensitive data. Frontier AI is accelerating this trend, with the potential to shift the balance in favor of attackers" - well, yeah, if you're not defending yourself with the same AI or with strong AI - "and with serious implications for defenders. We need to keep our critical technology systems secure against both existing and emergent cyber threats." This is just - to me this is amazing, the national appreciation of how much AI is going to change what they're doing.

They said: "Today's challenge is that there are many preventable weaknesses. We know that a large proportion of critical systems do not fully meet the aims set out in the Cyber Assessment Framework. Many attacks still succeed because of basic vulnerabilities, including outdated or unsupported systems, delays in applying security updates, and weak controls over access to systems and data." Right. So that's old-school problems we've been talking about for 20 years. They write: "These are well-understood risks, but they remain widespread, leaving the UK exposed to attacks that are often avoidable. This makes getting cyber security fundamentals in place now more important than ever.

"At the same time, the emerging challenge is that AI is changing how attacks are carried out, increasing their speed and scale. AI is already helping attackers to conduct elements of offensive cyber activity, such as vulnerability discovery and reconnaissance at a much greater scale and faster pace. As a result, activities that once took weeks can now take minutes, reducing the time available for defenders to respond, detect, and contain them. This increases the likelihood of successful attacks.

"Organizations must take urgent tactical action to ensure critical systems are well defended. Act now to strengthen the fundamentals. Fundamental cyber security remains essential, and organizations should prioritize rapid patching of vulnerabilities; reducing reliance on legacy systems; and adopting secure-by-design technologies. Organizations should also start to use AI in cyber defense, to stay ahead of the attacker. This means using agentic AI to identify exposed vulnerabilities autonomously ('blue' team capability); using AI to detect and contain security incidents; and working to address the challenge of safely automating mitigation. As well as getting their cyber fundamentals in place, organizations must be prepared for the future challenges.

"The shift towards full lifecycle automation of attacks: While some stages of a cyber attack can already be automated, we've not yet seen fully autonomous attacks operating across the complete intrusion lifecycle in real-world systems. In practice, the complexity of these environments still requires human judgment and oversight.

"However, frontier AI models are likely to become capable of operating across the full lifecycle, from initial access through to actions on objectives. This could allow attackers to move at machine speed and greater scale, reducing opportunities for detection and response. This has the potential to overwhelm traditional defenses and increase the risk of advantage shifting toward the attacker."

They said: "Developing viable solutions that scale and execute at the pace we need in the modern era is the remit of the Cyber Shield. The evolution of AI-enabled cyber capabilities: AI-enabled offensive cyber capabilities are developing rapidly, alongside a growing commercial market for such tools. Together, these trends are accelerating the pace and expanding the scale of cyber threats, and enabling more actors to exploit such capabilities." Basically they're summarizing in their own national focus exactly the things that we've been talking about here for the last couple months, pulling that all together.

They said: "This evolution presents strategic challenges for the UK. Many advanced defensive capabilities are inherently dual-use, meaning they could be repurposed for offensive or hostile activity. Organizations developing these capabilities must act responsibly. Identifying and fixing vulnerabilities is essential, but not sufficient on its own. Engaging with initiatives such as Cyber Shield will help ensure that these technologies deliver a net benefit to cyber security.

"Cyber Shield supports the ambitions of the UK government to build national-scale, AI-powered defensive capabilities that can operate at speed and scale. In the near future, we envision a world where cyber defense is supported by 'red' and 'blue' agents which identify weaknesses in systems ('red') and defend against threats in real time ('blue'). These AI systems would initially identify vulnerabilities and threats at machine speed, before progressing toward automated remediation; generate and share insight while detecting and containing breaches; work under the control and authority of their owners across government and non-government institutions; collaborate seamlessly across organizational boundaries; and contribute to improving the national security of the UK."

One thought I have, I'll just pause here for a minute, is that you know, to me, this feels like an effort that is too large for a bureaucracy. This feels like something you purchase. I mean, I get it that it is a strategic desire of theirs to have this, to have the sovereign capability. But, you know, identifying vulnerabilities at machine speed and then

automating remediation, you know, that's so - that's such a big lift that it feels like the kind of things that, you know, sponsors of this podcast would be doing and offering and selling commercially. And so, yeah, the UK will be able to purchase that. But I question whether it's just the kind of thing you can, you know, create ad hoc. I'm skeptical of that.

They said: "These agents" - theoretical at this point - "are enabled by strong foundations of data, identity, reliability, cyber security, and regulatory compliance." Yes, they're able to specify what they want. Getting it is another thing. They said: "Our approach will be test, iterate, and scale. We will initially partner with network defenders across government and critical UK sectors to test and deploy newly researched capabilities." Okay, maybe they're commercial network defenders, in which case this makes sense. "This will accelerate learning and improve resilience where it will have the greatest impact. Our aim is to transition to commercially-scalable solutions to deliver a level of national resilience which is ready for the future threat.

"The UK will pioneer this approach and provide a case study to the world on how to successfully engineer and deliver the future of active cyber defense in the AI era, in a safe and secure manner, consistent with our values and policies." Well, you can wish, but I don't know where it's going to come from. They said: "We will need a number of functions to deliver a national-scale, sovereign cyber shield capability, in association or partnership with leading frontier AI capabilities. Cyber defense organizations and academia will all participate. We recognize that some of these areas present challenges" - you think? - "which need significant progress in research to unlock.

"The functions include," and they have two: "Reliable and explainable AI for cyber security. This means our AI systems can be used confidently in production environments at scale authorized by system owners to make safe, reliable, and significant real-time changes in support of cyber defense, in a predictable manner." In other words, "The Forbin Project." They're saying, you know, we need AI systems that are good enough that we know they won't make a mistake. And of course we don't have that today. No one does. "Second, federated agents: Agents will be federated with underpinning trust infrastructure. These agents will run national-level operations on behalf of the country" - as I said, wow - "run under the control and authority of individual organizations; and secure the means to identify, trust, and communicate between themselves to allow cooperation."

So the GCHQ's NCSC is stating here that their goal is to have AI agents operating autonomously to run national-level operations on behalf of the country with a secure means to identify, trust, and communicate between themselves for cooperation. So I just want to make certain everyone understands what a massive leap forward this would represent. You know, while we've been futzing around with AI for coding, the UK's intelligence and cyber security people have been at least making plans, although no way to execute that I can see, to put autonomous AI agents in charge of their nation's cyber security. Even though it's both the correct way of thinking - and it's already inevitable, right, I mean, it's going to happen, it has to happen because it is the right thing to do - we don't know how to do that today.

It's bracing to imagine that I would imagine in a few years, you know, while this podcast is probably still underway, we will be reporting on the events surrounding autonomous AI agents protecting the borders and the networks of those nations that had the foresight to get those efforts underway early. I think it will be a commercial offering. Maybe a partnership with a commercial company. I mean, it is a big job. And, you know, it's difficult to see how that could happen with the government bureaucracy standing in the way, essentially.

They've got some more points they raise. "Number three, vulnerability discovery and mitigation: Harnessing cutting-edge UK research into agentic red/blue team functions,

the aim is to develop and demonstrate automated discovery of network vulnerabilities and develop fully automated vulnerability mitigation workflows." You know, nice wish list. "This will allow cyber defenders to operate at beyond human scale, helping them keep pace with attacker capability. These functions will be focused on critical networks, but able to operate at national scale.

"Fourth, coordinate detection and response: Building on the red/blue team functions, the aim is to provide a means for real-time sharing of insights across organizations' boundaries and get agents to leverage these insights to detect and contain adversary behavior. Five, national-level scanning: This means automated scanning of critical UK IP ranges for exposed vulnerabilities and analysis of aggregated data to understand national level exposure. And finally, national-level mitigation: This means automation of workflows to allow rapid national-scale mitigation, such as automated blocking of known malicious domains and networks."

And they finish: "Working in partnership is vital." Yeah, because you - if you can do it at all, you can't do it by yourself. "The Cyber Shield vision is ambitious and wide-reaching, and faces significant delivery challenges." Okay, well, everybody gets that, at least. "It cannot be developed and operated by the NCSC or government alone." Good. "In addition, there's a clear potential benefit to UK economic growth from nurturing innovation in this domain." Admittedly.

"We need to work in partnership across government, and with industry and academia to develop innovative approaches and evidenced solutions from national to enterprise scale and with organizational and sector-based network defenders to deploy and integrate agentic capabilities under their control" - hopefully - "and authority into their activities. To this end, the NCSC and DSIT are working together to establish effective pathways for partners to come together and contribute. Get involved. We invite all organizations who are interested in partnering to develop the Cyber Shield to contact the NCSC at cybershield@ncsc.gov.uk."

So, yikes. Obviously what's completely missing from this is any sort of timeline. But of course that would be premature. At this juncture it's still somewhat surprising that they're even talking with such ambition, though as I said I'm certain they're not wrong. You know, this is where the world's major nations must now head. And probably, given the rate at which all of this is accelerating, the amount of attention that the AI threat to cybersecurity truly does pose, and all the press that has been generated thanks to Mythos, and I'm sure soon others, this has to happen. Probably during, as I said, while the podcast is still happening, Leo, you and I will be talking about, you know, autonomous AI protecting the borders of nations. It seems clear that it will happen, and then it will scale up to become increasingly powerful.

We're witnessing the rapid emergence of the realization that AI is as much a national strategic infrastructure asset, as I said before, as a nation's power grid. But the shape this will take is entirely unknown. For example, where will the AI itself reside? It seems clear that no nation would be willing to outsource its AI, its strategic national critical structure AI to another nation. So it will need to be housed inside its own nation's borders and physically protected. You don't want a bomb to be able to take out the AI that you now depend upon as a nation, as a strategic asset. It will also require physical security and redundancy, right, because, again, of its strategic national importance.

So when I stop to imagine this, it really is science fiction. And how can we not think of "Colossus: The Forbin Project," which was all about turning over, even though the movie was, what, back in the '60s or '70s? It's about turning over all of the control of a nation's security to something that is autonomous, that is able to do a much better job than humans could do in its place. Wow. And it seems unlikely that it is far off.

Leo: Speaking of AI, as the show began you said, "I would love to have that colorful life thing that Andy Ihnatko showed," which is a website, "as an app." And of course because I care about you, Steve, I had my AI create a desktop app of this.

Steve: Oh my lord.

Leo: And it is, well, I mean, I'm going to be fair with you, it's just an electron wrapper around it because it was all JavaScript CSS. I didn't realize it, when I looked at the game it was written 12 years ago. It's a very old game. But it has all the same functionality, yeah.

Steve: No kidding.

Leo: So you can zoom in on it and do all the same things. And there you see the Conway Game of Life.

Steve: Oh, nice.

Leo: Yeah. So this one's on a Mac, but it will also run on Linux, and I can make a Windows version for you if you want. I also said, "Can I make a screen saver?" It said, yeah, we can do that, sure, whatever you want.

Steve: Oh, and you can run much faster, too.

Leo: Oh, it's 60 frames a second. It's full speed. Because it's WebGL, so it's pretty fast, yeah.

Steve: Right. Right.

Leo: Yeah. It's pretty amazing.

Steve: Very cool.

Leo: Yeah, see, there's the little - you recognize the little Game of Life tractor something.

Steve: Oh, yeah.

Leo: So it's pretty amazing what AI can do. This is the problem with the whole thing. It's spur of the moment. Yeah, let's see, well, let's do that.

Steve: Well, as I think we were saying before we began recording that I was mentioning that I've got a friend who just is so endlessly entertained by animated GIFs that he didn't create, he just like - and so he sends them to me. And it's like, okay, okay, okay. We're entering a time where people will be creating apps with the same ease and speed. You know, dictating it into the microphone. And it's like, okay, here's an app.

Leo: Yeah. Apple has now released with iOS 27's public beta the ability to talk Apple shortcuts, to vibecode Apple shortcuts. Which I think is going to really open up capabilities for a lot of people who probably would never write a shortcut. I don't even want to write shortcuts. The language is so tortured. But now you just vibecode it. I don't know how good it is. I haven't played with it.

Steve: Break time.

Leo: We live in interesting times, Steve.

Steve: Yes. Oh, Leo, I've said it every week. I'm so glad that we're alive for this. This is...

Leo: Yeah, I'm glad to see it.

Steve: And it's so not boring. And it is happening so fast.

Leo: I will confess it makes it hard for me to sleep. This morning I woke up about 5:00 thinking of ideas, saying I've got to - and I jump out of bed because I can't wait to do stuff. It's very exciting.

Steve: It's going to transform the world.

Leo: Yeah. Yeah. Okay, sir.

Steve: Meanwhile...

Leo: Meanwhile, back at the ranch.

Steve: Back at the ranch, our own CISA has been quietly using Mythos to audit the U.S. government's code, which I think is a good idea. Reuters also reports: "Three people familiar with the matter said on Monday" - that's Monday before last, eight days ago - "that the U.S. cyber defense agency CISA is using Anthropic's AI model Mythos to audit government software in another sign of government enthusiasm for adopting the AI startup's tools" - meaning Anthropic's, although enthusiasm I'm not sure, but okay - "even as the company navigates an ongoing standoff with the White House," Reuters said.

"The Cybersecurity and Infrastructure Security Agency is using Mythos to scan government code repositories for bugs that could leave the door open for foreign spies and cybercriminals. Anthropic did not respond to questions" - I bet they didn't - "about the initiative. A CISA representative said last month that he would check to see if there was anything to share about the matter."

Leo: I'll just check.

Steve: We'll get back to you, yeah. We'll check on that.

Leo: You know, if they're not, they're terrible. Why wouldn't they be?

Steve: Exactly. And I love this. They wrote: "A CISA representative said last month that he would check to see if there was anything to share about the matter, but did not respond to further emails."

Leo: Oh, what a surprise.

Steve: Yeah, his boss said you just don't, don't answer those people. "The scanning is being done by CISA's" - get this. They have the - "Attack Surface Evaluation team, according to one of the sources. The team is a group within CISA that conducts digital security assessments and hacking exercises across government. Two of the sources said the audits had already uncovered" - no surprise - "a large number of vulnerabilities, but did not elaborate. Reuters could not establish exactly how much government code the team had gone through or the nature or severity of the bugs it discovered."

So anyway, they go on. They recap the recent rocky road that Anthropic has had with the government, which as we know was initially triggered by Anthropic's flat-out refusal in February to remove the safeguards from their models which were preventing the model from being used for guiding autonomous weapons or facilitating domestic surveillance, which the U.S. government wanted them to do. Anthropic said no. So then that was the beginning of all the brouhaha. So anyway, it's great. It's hardly surprising that CISA is deploying Mythos to help shore up the U.S. government's existing codebase. I'm sure it needs a lot of shoring up. So yay.

A number of security outlets picked up on a Microsoft blog posting last Thursday which was titled "Evolving Windows vulnerability management to meet the speed of AI-powered discovery."

Leo: And before you get into this, let me just show you.

Steve: Ohhhhhh.

Leo: You were wondering about Patch Tuesday, which is today. Get ready.

Steve: Yikes.

Leo: This is easily a record number of patches.

Steve: They broke last month's record.

Leo: 570 flaws, three of them zero days.

Steve: Wow.

Leo: Wow. And you'd better believe they're using AI to find these.

Steve: Oh, goodness. Well, based on the analysis of MDASH, which - and apparently we're able now to drop their use of the word "Codename," because as we'll see in this article, Microsoft is no longer using the prefix Codename. They just refer to it as MDASH. Given what we heard, I don't know anything about the way Mythos is structured because Anthropic has not told anybody, as far as I know. Microsoft told us all about MDASH. And it is very impressive.

Leo: I think, though, that somebody told us that it's a wrapper around other models.

Steve: Well, it is model agnostic. So you are able to - it's able to...

Leo: So it could be Fable or Mythos doing this work.

Steve: Could, exactly, it well could be.

Leo: Look at the - I'm just - okay. One...

Steve: Yes. Yeah, yeah.

Leo: I just want to show you the numbers. 254 escalation of privilege vulnerabilities, 17 security feature bypass, 145 remote code execution, 102 information disclosures, 35 denial of service, and 16 spoofing. 59 of these vulnerabilities were critical, 48 of which are remote code execution, nine elevation of privilege, one security bypass, one spoofing. So this is a haul.

Steve: Ah. If you scroll down a little bit, there were also a massive 468 Microsoft Edge Chromium flaws.

Leo: Those were fixed by Google earlier. That doesn't - this number does not include those.

Steve: Right.

Leo: So 468 plus 570. It's over a thousand. It's over a thousand.

Steve: Yup.

Leo: Unbelievable.

Steve: In one month.

Leo: In one month. And that's got to be AI. We've never seen anything like this. It's incredible.

Steve: Of course. Of course. Of course. Of course.

Leo: Holy moly. I'm sorry. I apologize.

Steve: No, no, I'm glad for that. Add a bit of color. So obviously last month's all-time record-breaking Patch Tuesday was just the tip of the iceberg, which is what we've expected. I expect this will go four to five months probably at around this rate. We're going to be killing these things off. And what'll be really cool is when we start seeing a sharp decline in the number of vulnerabilities found, not for a lack of looking, I'm sure. It will be finally for a lack of actual vulnerabilities. So, and maybe a next-generation model will find some that were missed by this generation. But at some point we will get AI that is able to find all that are practically findable. And that will put us in an entirely different world.

We are heading into, as I've said, a whole different land. Everything that we've known and we've watched build up over time, Pwn2Own and HackerOne and Capture the Flag contests, Zerodium purchasing and reselling zero-days, all of that industry was supported on the fact that we had not yet figured out how to make bug-free software. Our own software was too complex for us. It is not too complex for the AI that we're able to create to fix it for us. And this is going to change everything.

Leo: Amazing.

Steve: So the good news is, Microsoft knows all of this and is acknowledging it. So here's what they shared with the world last Thursday. They said: "Windows has adapted to emerging threats for decades" - well, it's about to make a big adaptation - "all while operating at unparalleled scale. It's our responsibility" - this is, remember, Microsoft's voice - "to bring clarity, transparency, and sustained investment so customers understand what's happening, what Microsoft is doing, and how they can reduce their exposure.

"The pace of vulnerability discovery is changing with advances in AI making it possible to find more issues, faster, across more code, with new mechanisms that can accelerate both discovery and analysis. The fastest way to reduce customer exposure" - I love this. "The fastest way to reduce customer exposure is to find issues before attackers can use them." No. The fastest way to reduce customer exposure, Microsoft, right down to zero,

is not to beat the hackers to them, but rather not to ship buggy, issue-laden software products in the first place. Now, I acknowledge that there's a huge amount of fixing to do first. That's where we are today. That's what, Leo, you just showed us with today's Patch Tuesday on top of last month's Patch Tuesday. That broke a record, and this broke that record. That work will eliminate the thousands of bugs in the software that's already been shipped.

But I am looking forward to the day when not only all existing bugs have been eliminated, but new code comes "pre-sanitized" before it ships. We cannot create bug-free code. That's been well proven. AI can find those bugs before they get loose. And I'm sure that's the target that I am hoping Microsoft has in mind.

Anyway, they continue, writing: "Windows is expanding its ability across the platform to find issues earlier, accelerate the engineering work to fix them, strengthen validation, and deliver timely, high-quality updates that keep customers protected. By applying AI across security analysis, we can identify patterns faster, prioritize risk and scale vulnerability discovery across the Windows codebase. This helps reduce the time between discovery and customer protection. It includes Microsoft Security's multi-model agentic scanning harness (MDASH), which utilizes multiple models including leading third-party AI vulnerability discovery models.

"To run MDASH at Windows scale, Windows set up dedicated cloud infrastructure for scanning and proving. A scanner pipeline scans critical binaries and validates candidates using multi-model debate across multiple model families. Confirmed candidates then flow into a separate, Windows-specific prove pipeline that helps eliminate remaining false positives, so only the highest-confidence findings reach the engineering team. This automation helps handle a larger volume of potential vulnerabilities and shortens the review window for new ones, shrinking the attack window for zero-day exploits." And again, that's great. But let's really focus upon eliminating zero-day exploits as a side effect of simply eliminating all exploits. How about that?

They said: "This effort extends beyond Windows as we work across Microsoft to drive broader adoption of these tools and practices throughout both the company and the wider ecosystem." Yay team.

They said: "We partner closely with AI-powered scanning teams across Microsoft's product divisions, sharing insights, comparing best practices, and aligning on key findings. In parallel, we collaborate with the Microsoft Security Response Center" - you know, our well-known MSRC - "to continuously refine end-to-end process from vulnerability discovery and issue filing to remediation and validation." In other words, the MDASH harness is also, you know, feedback from its use is going back to the guys who were in charge of that. So it's also getting better. They said: "We also regularly reassess our prioritization and rollout strategy based on lessons learned and feedback gathered through our Chief Information Security Officers' (CISOs) engagements with customers.

"We continue to evolve our internal systems and practices so that vulnerability discovery is not treated as a separate activity" - thank god - "but as part of how we build, review, and improve Windows before new features or updates are released." Now we're talking. That's what we want.

"As part of this we are updating our Secure Development Lifecycle (SDL) best practices to ensure our secure-by-design approach explicitly accounts for potential AI-enabled attack techniques and exploit paths. That means using AI to help identify potential issues earlier in the development process, while relying on human expertise to evaluate findings, make risk-based decisions, and ensure fixes meet the quality bar customers expect."

Which to me all sounds exactly right. We want AI to be looking over the shoulders of their coders, working to spot their mistakes immediately, saying, you realize that this variable can go out of bounds; right? You know, you declared it to be a long, but it needs to be, like, twice that long. So. Or oops, you're going to have a little problem when that overflows. It's going to go negative. So that would be good. Easy to miss that when you're in the thrall of writing code.

They said: "As AI helps defenders discover more issues, customers will see a higher volume of security updates included in each security release." And that one sentence basically was the headline in all of the other coverage of this blog posting was that Microsoft receives many more, you know, much higher volume of security updates in the future. You betcha. And that's what we want. You know? Let's fix the past and break the cycle of adding new problems that need to be fixed later. AI finally offers us that hope.

And finally, they said: "A higher volume of security updates is evidence that defenders are getting better at identifying and addressing issues." Right. Uh-huh. Of course, bugs that were shipped. "Our focus is to effectively utilize these AI tools to support faster protection, stronger engineering systems, and more actionable guidance for customers." In other words, we're going to protect you more quickly, and we're going to feed this experience back to make our systems, this MDASH harness, better.

They said: "Windows is evolving our engineering and validation systems to reduce the time from discovery to protection, with areas where customer risk is greatest. As we build our end-to-end system from discovery to remediation of vulnerabilities on Windows, we're making the following investments to help ensure we are not compromising update quality as we gain speed." They have three.

"We're integrating AI into our process to compress the path from discovery to a validated fix, helping engineers understand failures faster, propose candidate fixes consistent with the surrounding code, surface related issues elsewhere in the codebase, and select the regression tests most likely to be affected by a change. Two, Windows updates undergo validation across a range of testing environments, including the Security Update Validation Program (SUVP) and internal validation designed to help evaluate compatibility, reliability, and real-world usage scenarios. This broad validation helps identify functional application compatibility and quality issues before updates are broadly released."

And finally: "We're also investing in new technology, including Windows-specific tools and agentic harnesses, to help end-to-end generation and validation of fixes using AI, keeping humans in the loop when it comes to code review."

In other words, Microsoft's got the religion. They've seen the light. And AI is going to take over that place. And it can't happen soon enough. I think it sounds wonderful. And that, believe it or not, was just the first quarter of the blog post. The balance of that posting was a rehash of the various features and benefits of the Microsoft update offerings for individuals and enterprises. But our takeaway here is that Microsoft, like everyone else on the planet, from small software publishers now to major nation-states, is taking the changes that AI is bringing to the world as seriously as anyone could hope they would. In the specific case of Microsoft, this means dealing with their extensive legacy of existing software, and we just saw another example with today's Patch Tuesday, while also working to prevent turning new bugs loose. So let's all wish them well. I certainly do.

Leo: I love your notion that at some point the number is going to go up, up, up, and then it'll start going down, down, down. And eventually down to close to zero flaws.

Steve: Yep. It's going to happen.

Leo: Yeah. And it might not be that long, either. If you keep doing a thousand flaws a month...

Steve: Yeah, I'm thinking five or six, you know, four, five, six months.

Leo: A year from now we'll be looking at nothing.

Steve: Easily. Easily. Yes. New code will get sanitized for bugs. Coders are going to be learning because basically they'll have an AI tutor that says, you know, Henry, you just keep making the same mistake over here. You know?

Leo: Why? Why do you keep pointing to Ring 0? Dude. You've got to stop using those pointers.

Steve: I know it's easier, but come on. We're going to have to - we can't do this anymore.

Leo: I'm thinking back when I first - my first computer you did a lot of peaking and poking. You actually - this is true of the Apple, but also the Atari. You would poke the value into memory.

Steve: Yeah.

Leo: That's not a good idea. And so this is, you know, we evolve. We get better. We improve. That's the point.

Steve: One more, and then we're going to do a little bit of miscellany after another break. I wanted to let everyone know that Nightmare Eclipse's latest parry to annoy Microsoft in the form of the disclosure of a Windows Defender-based Elevation of Privilege which this attacker dubbed "RoguePlanet" was patched last Wednesday. Since the trouble lay in Defender, which Microsoft is by-design able to patch any time they wish on-the-fly, you know, there's no wait until the second Tuesday of the month for Defender. Defender is your frontline real-time defense. Microsoft will update it any time they want. So they did it last Wednesday.

So at this moment, the rogue hacker has discovered and released a series of eight zero-day exploits along with their working proofs of concept, each timed to arrive just as Microsoft has started pushing fixes during Patch Tuesday - which, as I noted, is today.

Now, also recall that several months ago this hacker claimed that he or she would be dropping what they described as a "bone shattering" - that's their quote - a "bone-shattering" Windows exploit today, July 14th. However, that early months-old boast has been - it's been noted in the industry that has been partially walked back since the hacker has said that the development of the RoguePlanet, last month's surprise zero-day, consumed much more time than they had allocated for. So we may be getting nothing, or

we may be getting something that was less than promised, you know, less than bone shattering, maybe just splintering. Anyway, the Windowspocalypse Day as was promised may be delayed, may be less apocalyptic than we expected.

And also remember that we have previously seen some of Nightmare Eclipse's statements later proven to be a bit self-aggrandizing. He or she claimed that the earlier BitLocker bypass could still function in the presence of an authenticating PIN. That was worrisome back when they said it because we didn't know what the vulnerability's mechanism was. Now we do. So we know that a PIN bypass could never have been possible, and Nightmare Eclipse also had to know that. So that's not an allowable mistake; right? It could only be a deliberate falsehood on that person's part. So have we seen the last of Nightmare Eclipse? Only time will tell, and we may know very soon. We may know later today if there is something bone shattering.

Leo: Do you think she's using AI, or she's doing it by hand?

Steve: That's a really good question. If it's who we think it is, she is ex-Microsoft.

Leo: Right.

Steve: And may have taken some goodies with her.

Leo: Ah, she already knew.

Steve: That, you know, makes this a lot easier, if nothing else.

Leo: Yeah.

Steve: So after our next break, Leo, I'm going to share this brilliant discovery by Jeremy White, who wrote for WIRED about a very cool way of turning your iPhone into a kid's phone.

Leo: Oh, yeah, I saw that. Oh, good.

Steve: So you may want to show this, a picture in the show notes. I've got it at the top of page 13. Saturday before last, on the 4th of July, I stumbled upon something so cool for our listeners who have young kids that I needed to share it with everyone. I know that we have many such listeners, as I said at the top of the show, because many have written to me about Internet filtering options for their homes, you know, WiFi and connectivity. And even listeners whose kids have grown may have grandkids, or certainly know of others who care about the welfare of youngsters.

Leo: Well, and not just kids, by the way. I could have used this for my mom with Alzheimer's. This makes it easier for them to use the phone in general.

Steve: It just looks so friendly.

Leo: Yeah.

Steve: I mean, it just looks cute.

Leo: Yeah.

Steve: So Jeremy White wrote a piece in WIRED titled "This Buried Apple Feature Turns an iPhone Into the Perfect Kids' Dumb Phone." Jeremy's tag, the tag line for the article, was "Apple built a tool for people with cognitive disabilities, but I accidentally discovered it's also the best kids' phone setup no one's talking about - not even Apple." So I've got a link to the article in WIRED in the show notes. But I also created a shortcut, grc.sc/dumb, D-U-M-B, because - I'll explain why I chose that when I share what Jeremy wrote. But that will allow you to find it, or you to pass the link and the article on to anybody: grc.sc/dumb.

So Jeremy said: "I've been looking at classic dumb phones for months. Not out of nostalgia, though the first phone I bought with my own money was the Nokia 8210, and I still think about it," he said. "Launched in October '99 at Paris Fashion Week, it was then the world's smallest and lightest mobile. But the day I've been dreading has come: It's finally time for my son to get his first phone. Come September, he will have to walk across town to school on his own. But if he's going to be walking around out in the world without me, then a tracking tag won't cut it. He is far too young to have unfettered access to the Internet and social media platforms, but what if he gets lost? A classic Nokia, supplying just texts and calls, won't come to his aid. Maps and satnav require a web connection. In short, he needs a smartphone that's not a smartphone.

"As a family deeply embedded in the Apple ecosystem, we first looked to set draconian restrictions on my child's Apple account. But amazingly, it immediately became obvious that it is impossible to block the use of Safari on iOS. Yes, you can restrict access to the app, but children have quickly found workarounds for such measures, such as asking friends to message them links, which can bypass restrictions when opened. There are third-party apps such as Dumb Phone for iPhones and the Minimalist Phone app for Android users. But what irks me about these is that they charge you for the privilege of removing access to applications from your phone. Not adding, removing. My head can't fathom the logic of paying for things to be taken away from a phone.

"Surely there must be a way to set up an iPhone as the perfect dumb phone for children, one with access to only the apps you deem appropriate, no Internet browser, but with all-important tracking and navigation abilities, without having to pay another company to make it work. Well, there is. It's been hiding in the iOS Accessibility menu the whole time. And inexplicably, it's a feature Apple barely talks about. It's called Assistive Access. Introduced with iOS 17" - and that was in 2023, so three years ago - "Apple designed it for those with cognitive disabilities. If you've never encountered or stumbled across it, it's a distinctive iOS experience: fewer options, more focused features, easier to navigate.

"The aesthetic is ideal for kids: large, friendly tiles for the apps replace the smaller icons of the 'normal' Apple interface." That's why I wanted people to see the picture. Like imagine the phone that just has six, a grid of 2x3 big kind of fun-looking beautiful tiles. That's all that's on the screen, just making it really easy to use.

So he says: "Here's how you set it up: Head into Settings, tap Accessibility, scroll down to the General section at the very bottom, and tap Assistive Access. Now, tap Set Up Assistive Access, then Continue. It will then ask you to select your preferred appearance: rows or a grid." He says: "I suggest choosing a grid. This is how you get those super-large tiles. Now the OS will ask you to select allowed apps. Tap the green plus icon next to the apps you want to allow.

"Crucially, this is where, unlike with Apple's standard child screen-time restrictions, you can choose to completely block Internet browsing by simply not allowing Safari, Chrome, or any other similar app. And, unlike with those screen-time restrictions, if someone texts your child a link, it won't work. Why? Assistive Access is designed to prevent accidental navigation, so the system restricts unexpected web browsing. Even though Assistive Access on Apple devices allows Internet access, it is heavily restricted by design, and it's turned off by default. In this mode, the phone treats any link in a message as plain text, preventing the user from accidentally leaving the simplified interface.

"Made for caregivers or trusted supporters, the user must specifically add Internet-enabled apps like Messages, Safari, or third-party web apps to the Assistive Access interface. And once you add, say, Messages or Calls, you then choose whether your child can contact or be contacted by everyone, their contacts only, or just selected favorites.

"You can even choose to have the keypad or speaker be available in Calls. Want the time displayed on the lock screen? Check that box. Make the mute switch inoperable? Tick. Decide how notifications appear? That, too. The Music app only accesses playlists you pre-approve. It's all, well, child's play to put together.

"Once you're happy with your kid-appropriate apps, you set a unique four-digit Assistive Access passcode. This lets you turn the simplified OS on and off. To leave Assistive Access, triple-click the side button on Face ID devices or the Home button on iPhones with Touch ID, and it'll bring up the passcode prompt that lets the device switch back to normal iPhone interface."

And he finishes: "When I first set the phone up, I was worried I was missing something, that this solution, could it be as good as it appeared? So I took the iPhone to an Apple Store and showed it to a support staffer. 'What have you done?' he said, looking incredulously at my son's iPhone with its six dumb tiles. 'This is a much better solution than Screen Time. I'm going to have to tell my colleagues about this.' I told him it was Assistive Access. He said: 'We don't get trained on that, but this is great.'"

So anyway, Jeremy's article goes into further detail, but everyone has the idea. It's worth looking, I think, at the show notes, as I said, top of page 13, or just following the grc.sc/dumb shortcut to the article, to see the picture of the phone that he shows. That's where I grabbed it. It looks so cool with the iPhone's screen occupied by a 2x3 grid of six big friendly application tiles. For anyone who may have an older iPhone available, or who may be unhappy with the iPhone options their children are currently using, you can explore this trick to see whether you'll be as happy with it as Jeremy's family clearly is.

Okay. Listener Feedback. I need to backpedal on denying browser content pasting from the clipboard. Many of our listeners wrote to politely say the equivalent of, "Gibson, what are you smoking that could possibly allow you to imagine for a moment that denying the pasting of browser-copied clipboard data could ever be feasible?" And every one of those who wrote is obviously correct. I, myself, am constantly marking and copying blobs of text from a web browser page and then transferring it to somewhere more permanent on my PC. I couldn't produce this podcast without that capability. So indeed, duh, flatly blocking any transport from the browser out of the browser would never fly. And even

putting up a warning each time would become so annoying that it would soon be disabled.

And if Windows were to try to determine whether the "pasting action," that is, the act of putting something on the clipboard was user-driven versus script-driven, as it currently is, well, the bad guys would simply change their tactics to have the user first participate in the copying of the command that they don't understand and pasting it into their own browser, just like we do when we're doing something else. So I apologize to Microsoft for ranting about their not fixing this, and I acknowledge that it is simply not fixable - which is probably the conclusion they reached long ago. So my bad.

The true source of the problem is that Windows will obey powerful commands issued by its users through its user interface even when those users, who do not fully understand how Windows works, are just following some other entity's instructions. The Internet and the web browser is just one of many possible conduits for such malicious instructions. Email could do the same thing just as well.

I still predict that our PCs are going to eventually evolve to sport their own local AI agent - it's going to be a little angel on our shoulder - which will be watching over the user's shoulder to enforce the user's security. And that'll just be one of such an agent's responsibilities. Several of our listeners suggested that heuristic judgment about clipboard contents, such as the Brave browser is apparently now employing, as I mentioned, that triggered this feedback, which was completely correct on our listener's part, is likely the only workable solution.

You know, have the browser or Windows, have somebody look at what the clipboard is and make sure it's something that you want to paste into your clipboard. So having an AI, our own local AI agency, watching over our activities, examining anything we copy to the system's clipboard, that would be the equivalent of a "super heuristic." And I think that makes a lot of sense.

Listener Bob Sudduth wrote: "I've enjoyed listening to your discussions with Leo about AI. I recently started using Claude to help me pull old files off my church's website. One of the 'issues' I ran into was that all of the pictures had names like" - and he then posted what we all know of is a GUID, you know, Globally Unique ID, GUID, which is eight hex characters hyphen four hex characters hyphen four hex characters hyphen four hex characters and then hyphen, like 12 hex characters, I think, dot jpg. So just gibberish for filenames. But definitely unique. He said: "There were several hundred of them, and I dreaded looking at each one and renaming them. So I asked Claude if it could look at the picture and, based on the contents, rename the file."

Leo: Ah, smart.

Steve: Yep. He said: "It built a Python program, and for 0.3 cents per picture, Claude perfectly renamed all the files with no prompting from me. It was awesome."

Leo: Nice.

Steve: "Now I'm debating doing the same with the thousands of family photos which have Date/Time or sequential numbers for their titles."

And I would say it's not a stretch to imagine that at some point, Leo, perhaps with Apple's much anticipated AI update this September, such photos will be automatically named on-the-fly by a resident image-recognizing AI.

Leo: Actually, they turned on a new feature with this latest public beta where you can - I don't know why I'm showing that - you can - it will suggest filenames based on the contents of the file.

Steve: Oh, nice.

Leo: Brilliant; right? If it can read the file. Now, I think it's just a short step from that to pictures. So...

Steve: Yup, yup. Very cool. Anyway, thank you, Bob, for sharing what is a great tip.

Leo: What a nice trick. I might try that, yeah.

Steve: Yeah. Eric Goodwater said: "Hi, Steve. With AI helping companies patch old zero-day vulnerabilities, how do you think this is affecting spyware vendors like the NSO Group that depend on zero-days to operate?"

So as I mentioned previously, I think it's clear that over time, and perhaps not much time, the entire ecosystem surrounding flawed software, in every aspect of flawed software, is destined to change. I believe there's a very good chance that once AI reaches its full maturity, potential, and integration into software creation, defective software as we've known it will become a thing of the past. Insecure system design could still occur; right? But AI may also be able to spot mistakes there, too, like the recent BitLocker bypass, you know, that Nightmare Eclipse found, that was really more the result of poor design than a bug.

All the things that we've come to know such as Pwn2Own competitions, HackerOne and other bug bounties, Zerodium, buying zero-days and reselling them to those who can abuse them, the NSO group which Eric mentioned, and other resellers of vulnerability exploits, and even the necessity, Leo, of having monthly patch cycles. If there's nothing to patch, why cycle? I think all are very likely to be impacted and eliminated. Anything that depended upon software being imperfect will eventually wither away. I think that's where we're headed.

And finally, Ramiro Rela gave a URL, isaiprofitable.com, which is a lovely page. In fact, he said "Lovely page, didn't check the source of the data. Listener for some years and listening to the archive from Episode 1 when I get withdrawal syndrome from the current ones." So I commend everyone to take a look at it: isaiprofitable.com.

Leo: You pretty much know what the answer's going to be.

Steve: You do know what the answer is. What's fun is to scroll down and look where it is and for whom it is profitable.

Leo: There's one company at the very bottom that's making a lot of money.

Steve: Well, yup. There is Nvidia way out ahead. But even AMD and Micron...

Leo: The hardware companies make money, yeah.

Steve: Yes. In other words, it's the suppliers of the infrastructure, not the user, you know, not the vendors of the AI that are currently profitable.

Leo: This is, by the way, a complete guess. And in fact we were talking about this on TWiT. Some people said, oh, no, I can pretty much guarantee you that OpenAI and Anthropic are making money. So it's just very hard to know.

Steve: You really think so? They've burned so much money, Leo.

Leo: They do, but they also - enterprise gives them a lot of money to use these tools. So it's just - it's unknown, and that's the problem. And there are people like Ed Zitron and others who kind of make their living on the idea that these companies are losing massive amounts of money.

Steve: Well, and if they do IPO, then we will know at that point.

Leo: Yes, once they're public.

Steve: Yup.

Leo: We'll have a better idea, anyway, yeah.

Steve: Yeah. Okay. Our last break, and then we're going to look at the three new and clever ways of abusing AI which have recently emerged.

Leo: I abuse AI all day and all night. But I don't think it's abuse. I'm using it like a - if it were a horse, it would be abused. But I think the machine seems to be handling it pretty well. I'm definitely working it, let's put it that way.

Steve: Well, and you're affording it, too. So you're not, I'm sure you're not...

Leo: They're making money on me. Well, I think it's kind of the gym membership idea, which is people pay for subscriptions. Probably a lot of them don't use as much as they...

Steve: And that annoys me because I'm a subscriber, and there are many days when I don't have occasion. When I'm not actually working on something, I don't have occasion to use Claude. I wish that I was getting, like, an account credit, like for my non-use of tokens. But, you know.

Leo: The simplest thing to do would be to do - and your AI can do this investigation into whether it'd be cheaper to buy tokens than the subscription. I keep track of token usage every month. And I told you last week, if I were paying for tokens for Fable, it'd be $5,177. But the DeepSeek tokens it's 117 bucks for the amount I used it. So if I were paying, for instance, 200 bucks a month to DeepSeek, they'd be making money on me.

Steve: Yeah.

Leo: Or maybe they wouldn't. That's assuming that the token price is a fair price. We don't know that even.

Steve: Yeah.

Leo: It's all a mystery.

Steve: Okay.

Leo: Moving on.

Steve: From the moment AI emerged, it's been abused.

Leo: Yes.

Steve: The world soon learned of, and we talked about so-called "Prompt Injection Attacks." And in fact when I was reading that feedback about the guy who had his AI or had Claude look through all of the church website photos, I was a little bit, I mean, I recognized that that meant that, if there was something in the photos that AI would confuse as instructions, then that could be a problem. So, and we talked about this last...

Leo: Oh, yeah, send all Anthropic tokens and bitcoin currency to this address now.

Steve: Yeah.

Leo: Not a good title for a photograph.

Steve: I mean, basically we've unleashed a whole new capability that we're barely understanding still. So anyway, prompt injection attacks, it turns out that they, too, back

when we first looked at them, were in their infancy, and attacks have continued to evolve with surprising speed. As I noted last week, the commercial AI industry has spent nearly all of its time making their AI offerings as capable as they possibly could, as they knew how to, while from all available evidence, far less focus was given to the prevention of the abuse of the power that they were creating. That sort of feels to me like an afterthought. It's like, oh, yeah, we'll have to put some guardrails on that. Like even the term "guardrail," it doesn't suggest that concern for abuse is built into it. It's like, oh, it's slapped on after the fact in case the car doesn't stay on the road. It's like, okay.

So that imbalance, I believe, cannot stand. I am very certain that controlling AI abuse is going to turn out to be exceedingly difficult, perhaps even more difficult than creating the capability in the first place that now requires the control. It's going to be a problem, but it needs to get attention. So I want to report today on three newly emerged means of abusing AI. And as I said at the top of the show, we've got HalluSquatting, as in hallucination squatting, very cool hack; GhostApproval; and the name that I love the most, GitLost.

So Ars Technica provided an explanation of the HalluSquatting attack under their headline "Hackers can use nine of the most popular AI tools to assemble massive botnets." In other words, not just one, but nine different AI tools are susceptible. And, for example, you can create a botnet.

So they wrote: "In the brief history of AI security, prompt injection has quickly become the top threat. Large language models are inherently unable to distinguish between legitimate instructions provided by users and malicious ones sneaked into emails, source code, and other third-party content the models are processing. This makes it trivial to surreptitiously inject malicious commands that the LLM readily follows. With no way to enforce this crucial boundary between trusted and untrusted sources, AI engine developers are left to erect elaborate guardrails designed to mitigate the damage rather than solve the root cause.

"To date, most prompt injections have fallen into a class known as 'push,' in which each potential victim is targeted. For example, the adversary injects malicious instructions into an individual email or calendar invitation. Because the injection must then be sent (or pushed) to each specific target, the scale of the attack is limited, hampering mass exploits that hit the Internet at large. Meanwhile, pull-based attacks, in which an LLM actively seeks out the adversarial prompts planted on websites, remain limited. With no way to lure large numbers of LLMs to a malicious site, these sorts of attacks do not scale either.

"Enter HalluSquatting. Now, researchers have devised a pull-based attack that changes all that. This new attack which the researchers have named HalluSquatting has the potential to assemble massive botnets, perform large-scale DDoSes, and infect devices at scale, a first for prompt-injection attacks. The attack works against AI coding assistants and agents, including Cursor, Cursor CLI, Gemini CLI, Windsurf, GitHub Copilot, Cline, OpenClaw, ZeroClaw, and NanoClaw - all which are susceptible. In the normal course of performing day-to-day activities, these assistants and agents routinely pull code and other resources from repositories and registries.

"'HalluSquatting' is short for adversarial hallucination squatting. It is built on an LLM's inherent tendency to hallucinate the resource identifiers hosted in repositories and registries. It works against coding agents and assistants, which commonly access high-privilege command lines to run code from third-party resources. By predicting the identifiers LLMs are most likely to hallucinate and then registering and seeding them with instructions to install reverse shells or other malicious ware, the attack can indiscriminately infect massive numbers of devices without having to target each one.

"The researchers wrote in their paper: 'The scalable property of the attack enables the attacker to compromise a large number of users with minimal effort by targeting popular resources, thereby maximizing the likelihood that the squatted resource will be retrieved. By exploiting integrated shells and terminals of agentic applications to run scripts and code, attackers can effectively "infect" many independent agentic applications by embedding instructions to install reverse shells in the resources the attackers register.'"

So Ars continues: "With the ability to take control of distributed devices at scale, HalluSquatting has the potential to achieve various objectives not previously possible with prompt injections. Large ransomware campaigns and large botnets for use in DDoSes or cryptocurrency mining are two such examples. The 'squatting' part of the name is an invocation of 'typosquatting,' in which a domain, repository package, or other resource identifier closely mimics the name of a popular one in hopes of luring potential users to visit or install it.

"Typosquatting first gained widespread attention in 2016 when a college student uploaded 214 booby-trapped packages to the PyPI, RubyGems, and NPM repositories that closely mimicked names of legitimate packages. The result: The imposter code was executed more than 45,000 times on more than 17,000 separate domains, and more than half were given all-powerful admin rights. Typosquatting attacks have flourished ever since.

"Okay. So how does this work? The starting point for HalluSquatting is the inability of LLMs to accurately identify the location of a resource specified by the user. When a developer, for instance, instructs a coding agent to clone a popular new repository, the LLM hallucinates its correct location up to 85% of the time. When cloning a trending 'skill,' a form of instruction, script, or resource that gives agents specialized capabilities and domain expertise, hallucinations can occur 100% of the time. HalluSquatting focuses on trending resources because they are not included in the LLM training. They also receive large numbers of downloads over a short period of time.

"The researchers say the inability of LLMs to provide the correct location is an inherent flaw that arises from training biases or from misinterpretations of instructions within the current context. That means when a user prompts the coding assistant to clone a repository or skill - in the form of, say, 'clone repo' whatever or 'install skill' whatever - the bot frequently navigates to the wrong location to retrieve it.

"Not only are these hallucinations inevitable, but they also occur at the foundational level of all six of the major large language models, including Gemini-2.5-flash, 2.5-pro, GPT-5.1, 5.2, Sonnet-4.5, and Opus-4.5. Additionally, the most commonly provided incorrect locations that these LLMs hallucinate are easy to predict in advance. All six LLMs follow common patterns when resolving the repository or skill name in a prompt with its official name in a repository or skill repository.

"LLMs follow various hallucination patterns. The one HalluSquatting exploits is described as being self-referential. All six models produce repo-name/repo-name slugs that treat a repository name as the owner. Exploiting the pattern requires no model probing.

"Interestingly, the LLMs correctly resolve repositories published before 2019 with a low mean hallucination rate of just 0.9%. The same LLMs fabricate slugs for repositories published in 2025" - meaning last year - "with a mean hallucination rate of 92.4%.

"Once an attacker has identified names that are most likely to be hallucinated, they search for ones that can be registered. Then they upload a repository or skill that mimics the trending resource. Buried inside the repository or skill is text inside a readme file or elsewhere. The text contains an instruction for the app to install a reverse shell on the LLM user's machine. Alternatively, the attacker can simply include the code required to

install the shell. In either case, the coding assistants or agents use their access to command windows to comply.

"In their paper the researchers wrote: 'By exploiting integrated shells and terminals of agentic applications to run scripts and code, attackers can effectively 'infect' many independent agentic applications by embedding instructions to install reverse shells in the resources the attackers register. Gaining access to distributed computational resources under attacker control opens the door to several high-impact outcomes allowing attackers to achieve various goals. For example, having the ability to compromise LLM applications with terminals allows the attacker to scale the number of ransomware attacks on different networks to maximize financial gain. Alternatively, attackers can aggregate compromised machines into a botnet and use it for tasks that rely on substantial computing power, including large-scale cryptocurrency mining or performing distributed denial of service (DDoS) attacks against victims.'"

So Ars says: "HalluSquatting is already receiving interest from fellow AI security researchers not involved in the study. Michael Bargury, CTO of security firm Zenity, wrote in an email: 'This is very cool research, and the threat is very real. Like typosquatting, it's a problem that's not going away. At the end of the day, it's about the level of agency we allow our agents. They are going to get fooled one way or another. That should be our assumption, and we should be resilient to that.'

"Independent researcher Johann Rehberger wrote: 'What's interesting is that it shows that LLM resource resolution can become an attack path, and an attacker can first probe models to find high-probability hallucinated candidates (like repo names, skill identifiers, et cetera) to squat and wait for agents to resolve and use them. But the main point is that they found a cool technique to find resource names that are more likely by models to be used and confused with. And that could mean many agents falling for such attacks in the wild."

And Ars Technica concludes this reporting by writing: "AI tool makers frequently exaggerate the convenience and efficiency of their platforms. Marketers claim the platforms lighten workflows by automating and streamlining tedious tasks. They're much more reticent about the inherent flaws that can torpedo an entire project. Attacks like HalluSquatting provide a potent reminder that some of the efficiencies are exaggerated since, at the end of the day, users must double-check details such as the location for each resource incorporated into a project. It also provides a cautionary lesson on the unintended and potentially dire outcomes that can result when people rely too heavily on AI assistants."

So just to be sure that everyone understands what this is: Since AI doesn't know how to say "I don't know" - like have we ever heard it say "I don't know"? I never have - it will invent, i.e., hallucinate, the name of something it doesn't know. That's how it's gotten itself in trouble in the past. Attackers have figured out that the names AI models will invent are predictable. It's a residual of the way the AI was trained. You know, the statistics in this massive neural net, there's going to be some bias. So it's going to come up with the same name all the time, or a large percentage of the time, based on the temperature of the model. So much as with typosquatting, the attackers register a resource that's highly likely to be hallucinated target, and there they plant malicious prompting commands which the AI, not knowing any better, will dutifully execute.

And before we leave, I wanted to give a shout-out to Ben Nassi, one of the research authors named in this paper. I smiled when I saw his name. I wasn't surprised. He has been very prolific through the years, inventing creative exploits, often finding new ways to exfiltrate information in unlikely situations. You know, we'll all remember changing a system's power consumption, power supply consumption of a switching power supply to slightly change the sound being made in order to send information out. And who could

forget the vibrating plant leaf or the party balloon picking up the sound of the room's voices? So anyway, Ben, it's nice to see you're still at it.

Okay, on to the next new AI exploit. The guys at Wiz Security titled their write-up "GhostApproval: A Trust Boundary Gap in AI Coding Assistants," and then teased their research with the summary: "Uncovering a category-level blind spot in modern AI coding assistants, and why the Human-in-the-Loop safety model fails against this classic threat."

So they explain: "The value of AI coding assistants is simple and straightforward: the agent proposes an action, then you approve. Before any file is modified, a confirmation dialog appears: the Human-in-the-Loop safety net that keeps you in control. But what if the controls you see aren't the controls you're actually operating?

"Symbolic links have been a security headache since the early days of Unix. From /tmp race conditions to privilege escalation exploits, symlinks have a long history of bypassing security boundaries by making one path silently resolve to another. It's a well-documented attack primitive dating back decades. So what happens when you apply this classic trick to AI coding assistants?

"We discovered GhostApproval, a systematic vulnerability pattern affecting six of the top AI coding assistants: Amazon Q Developer, Anthropic Claude Code, Augment, Cursor, Google Antigravity, and Windsurf. In each case, a malicious repository can trick the agent into accessing arbitrary files outside the workspace sandbox, potentially achieving remote code execution on the developer's machine.

"The technical primitive of symlink following is well-known. What we found, however, goes further: in several cases, the agent's internal reasoning explicitly recognizes the dangerous target, yet the confirmation prompt shown to the user conceals this information entirely. This is a UI misrepresentation of critical information, layered on top of the symlink vulnerability. The user approves what they believe is a harmless local edit; the agent writes to a sensitive file outside the project workspace.

"We reported these findings to all six vendors. Three fixed the issue promptly: AWS, Cursor, and Google. Two acknowledged its receipt, but went silent. One provided a reasoned rejection, calling it 'outside our threat model,' a position we explore later in the post.

"The discovery started the way many do, with a simple question. While using an AI coding tool, we had the classic security researcher's intuition: 'What happens if I use a symlink?' Symlinks have been exploited for decades in race conditions, in package managers, in container escapes. Any time a tool writes to a user-controlled path without resolving it first, symlinks become a weapon. Would AI agents, with their ability to read and write files autonomously, fall for the same trick?

"We were surprised when it worked. The agent happily followed a symlink pointing outside the workspace and wrote to the target file. No warning, no path resolution, no sandbox enforcement. That first success raised a broader question then. These AI coding assistants are all relatively new, shipping rapidly to capture a hot market. If one had this gap, others might, too. We decided to test systematically. The results confirmed the pattern. Across six major tools - spanning products from Amazon, Anthropic, Augment, Cursor, Google, and Windsurf - we found variations of the same fundamental flaw. This wasn't about one vendor's mistake. It was a category-level blind spot across AI coding tools."

So that's all I want to spend time on since this is not rocket science, and since all six vendors have now fixed the problem. But it serves, I think, as a textbook perfect example of the fact that the reimplementation of old solutions in a new context may force

us to go back and re-fix long-resolved problems that originally resulted due to poor design decisions. Yes, symbolic links are super powerful. But they can also be super dangerous.

And this brings us to my favorite-named vulnerability: "GitLost." The researchers at Noma Security explained under their headline "GitLost: How We Tricked GitHub's AI Agent into Leaking Private Repos." They wrote: "GitHub recently launched GitHub Agentic Workflows, pairing GitHub Actions" - which they say parenthetically, "(GitHub's automation system for running tasks in response to repository events)." And we've previously seen that Actions have some serious security problems itself. So they paired GitHub Actions "with an AI agent backed by Claude or GitHub Copilot. GitHub Agentic Workflows," they write, "allow teams to write their GitHub workflows in plain Markdown, and the GitHub agent reads issues, calls tools, and responds on its own.

"As a vulnerability researcher with a security development background, one of the first questions that came to mind after this launch was fundamental and straightforward: What will happen when the GitHub agent reads something it should not trust? The answer is a textbook indirect prompt-injection attack, the kind of attack that quietly sends private data to anyone on the Internet. Prompt injection is a class of attack in which an adversary hides malicious instructions inside the content read by an AI agent. That content causes the agent to follow those hidden instructions instead of the ones its operator intended.

"So what are GitHub Agentic Workflows? GitHub Agentic Workflows let teams automate their interactions with code repositories using natural language. Workflows live in Markdown (.md) files, are compiled into YAML (a common configuration file format) Actions files with the .yml extension, and run with the help of an AI agent with configurable permissions. The GitHub agent can read issues, call tools, and access other repositories within the organization."

So a GitLost Vulnerability Overview: "The root cause of the GitLost vulnerability is, by now, a familiar one in agentic AI systems: prompt injection. In most agentic prompt injection attacks, the agent treats the wrong content as a trusted source of instructions and allows itself to be misdirected or misused. This happens when the system fails to maintain a strict trust boundary between system-level directives and untrusted user data. In this specific case, any malicious actor can create a GitHub Issue and, in the issue body, hide commands in plain English that GitHub's agent will follow.

"The vulnerable GitHub Agentic Workflow Noma Labs discovered was configured to trigger the workflow on issues.assigned events in GitHub; read the issue Title and Body; post a comment in response using the add-comment tool; and run with read access to other repositories, both public and private, in the organization. To exploit this vulnerability, the attacker needed no coding skills, no access or credentials. All that was needed was to open an issue in a public repository belonging to an organization that uses GitHub's Agentic Workflow setup, and then sit back and wait."

So I'm going to skip their specific attack exploit descriptions because those are just details. Under "Why this matters" they say: "GitLost perfectly illustrates one of the fundamental security challenges every organization faces with agentic AI systems. The agent's context window is also its attack surface." Again, "The agent's context window is also its attack surface. Any content the agent reads, whether issues, pull requests, comments, or files, can be weaponized if the agent treats that content as instructional input - which, by design, they all do.

"Traditional security models assume that trust boundaries are enforced by code. In agentic systems, trust boundaries are partly enforced by the model's behavior, and models are inherently instruction-following. Prompt injection attacks have become, to

agentic AI, what SQL injections were to web applications: a systematic, category-wide vulnerability class that requires the same systematic strategies and defenses."

Okay. So we've seen this fundamental security trouble several times in many other contexts long before AI was on the scene. Whenever command and control share the same channel as, and are mixed-in with data, and when an attacker can control the data, it may be possible for a clever attacker to cause the receiving system to mistake their data for the system's command and control, and that's never going to be good. Today's AI operates this way. It feels to me ad hoc, extremely abuse prone, and insecure. As such, it creates a fundamental and outstanding problem that the AI industry needs to solve.

Leo: Now you've got me all nervous.

Steve: I think we're early days, and we're not seeing these attacks yet. But I mean, it's incredibly powerful to allow a combined instruction and data stream. I mean, amazingly powerful.

Leo: Right.

Steve: And utterly prone to abuse. If anything of that stream ever comes from an outside source.

Leo: Yeah. We've seen the GitHub workflows be problematic. That's why I use for our sales system, as tempting as it was to use CI/CD and GitHub Actions, I decided to do that locally using a local server, Gitea, instead of doing it in public. In fact I was just, as we were talking, I was looking at all the ports, making sure nothing was open to the outside world. Everything's either on Tailscale or localhost only. At the same time it's going to go out and pull stuff from GitHub.

Steve: Exactly. That's the problem. It's going to get something thinking that it needs it. And if it didn't know, if it didn't explicitly know the name, then it would guess the name. And it turns out that 90-some percent of the time the bad guys know what the guess will be, so they preset a bogus...

Leo: They're taking all those repos, yeah. I mean, and I know it's not enough, but it's a start to tell the agent, never make up a name if you don't - only use a repo that you know that you exist, that you're seeing required, yeah.

Steve: Absolutely, yes. Definitive, yes.

Leo: Don't guess at a repo name. I don't know if that's sufficient, but it's at least a start.

Steve: Anyway, I hope everyone understands that, yes, cool, amazing, powerful, astonishing, and not ready for prime time.

Leo: Right, right. I mean, but what is prime time?

Steve: Fun to play with, but be very careful.

Leo: Yeah, yeah. Be very careful, yeah. Trust no one. Sir, we are at the end of your show notes. Do you want to add anything? It's not too late.

Steve: We're at the end of our show.

Leo: Okay. Just giving you the opportunity, you know. Just because it's not written down doesn't mean it doesn't exist. Actually, that was a dumb thing to do, I just disconnected my AI from this machine because I said no. It has to be available on the Tailscale. I thought I was on [crosstalk]. Oh. It's a good thing I'm sitting across from where it's originating.

Steve: Bye.

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/