Meta Patch Tuesday Fallout

Description: Andrew Ng weighs in on AI Doomsaying. The wisdom of outsourcing AI security testing. The true risk of an AI-created bioweapon. The EU KIDS Act - this one is even messier. "Nightmare Eclipse" finally unmasks himself. A white hat firm used Claude to attack OpenAI. Cisco's own massive 77 CVE update. What was the fallout from September's Patch Tuesday?

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1097.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1097-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. Lots to talk about. Amazing problems that cropped up after Microsoft's Patch Tuesday. Steve will run through those. We'll also talk a little bit about the AI doomers and why Steve's not at all worried. He also has his thoughts about the EU KIDS Act, which I suspect you already know, but you might want to listen. Security Now! is coming up next.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1097, recorded Tuesday, September 22nd, 2026: "Mega Patch Tuesday Fallout."

It's time for Security Now!. Yay, Tuesday is here. And so is Mr. Steve "Tiberius" Gibson. Do you think when you're 95, Steve, you'll be doing heavy metal concerts like William Shatner is doing now?

Steve Gibson: He is amazing.

Leo: He's wild.

Steve: He is. I just think it's great. And I love that when someone says what is your secret to longevity, like what can you tell us, and he said, "Don't die."

Leo: Yeah, that's the best advice ever. Steve, what are we covering on today's mega episode?

Steve: So, yes. For Security Now! 1097, this is the penultimate September episode.

Leo: He just likes to use that word, kids.

Steve: Now I know what it means. I want to spend some time talking about the consequences of Microsoft's Mega Patch Tuesday. We touched on last week, oh, well, they broke copy and paste for Excel. Which you didn't think was a big deal. Of course all the people like whose workday is Excel spreadsheets, they were quite discomfited by that. Turns out that wasn't even the tip of the iceberg of what happened. So it's interesting that the good news is that nearly a thousand big problems -

there were, what, 119 or 117 critical problems solved a couple weeks ago? Those are gone from Windows. Turns out, though, that there was some fallout from all of that, which we're going to talk about.

But first I wanted to pull in some expert opinion about all this AI doomsaying. Andrew Ng weighed in just a couple days ago. Of course he's somebody who knows something about AI. I want to share his thoughts about that. Also he referred to an individual who actually knows something about the biorisks of AI. So we're going to hear from somebody who's actually been there, who is actually both an AI expert and somebody who's made viruses so is able to cross over as one would have to. Also, you and I were talking about this before, but it's very interesting that all of these breakouts had a single locus that nobody else, I haven't seen anybody else in the industry talk about this.

Leo: I mentioned it on Sunday, I'm proud to say.

Steve: Good. Because we're going to talk about what I call the "wisdom of outsourcing AI security testing." I'll bet everyone's going to stop that in the future. Also we've got the EU trying to push their KIDS Act. And boy, this one is even messier than the things they've tried before. Nightmare Eclipse has finally unmasked himself. We know the back story behind that, which we'll take a look at. Also a white hat firm, a security firm, used Claude to successfully attack OpenAI. There are some very interesting lessons there we're going to get to. And then Cisco is clearly on the AI train now because they just pushed out what is for them a massive 77 CVE update. And oh, my god, they're all 10.0s and 9.8s, I mean, it is - I hope everybody who's using any Cisco equipment anywhere is patching this because of course the bad guys are going to jump on that because it's such a big...

Leo: Can you believe Cisco? Geez.

Steve: Well, they're another Microsoft; right? They had all these problems, and they seemed unable to fix them. But what they do have is what Microsoft has, which is cash. And so if you can purchase fixes for your legacy software blunders, which is what Microsoft has done with AI, and what Cisco is now doing, they're just buying the fixes by paying AI to do it. That's good news that they're doing it because they're going to end up with safer systems than they've had before. And then we're going to look at the fallout from this Patch Tuesday, which was quite significant. So I think lots of interesting stuff for our listeners. And of course we've got a Picture of the Week that, if you haven't seen it, Leo, several people sent it to me. It may have made the rounds. Not sure. But it is kind of wonderful. So we've going to have, I think, a great podcast.

Leo: I will let you know when I look at it because, as you know, when I get the show notes, I avert my eyes.

Steve: You try, you endeavor to expose yourself...

Leo: I don't want to see page two. So yes, the Picture of the Week coming up, as well as a, as usual, stellar episode of Security Now!. So glad you're here, Steve. So glad you're all here. Now let's share the Picture of the Week.

Steve: So I gave this the caption, it really doesn't need one, but I said: "The more experience one has with coding and people, the more this price list makes sense.

Leo: Ah, price list. And this is on one of those boards that you put outside your office; right? Go ahead, Steve, tell us what this says. It's real, too, by the way. I can see the photographer's reflection in it. It's so great.

Steve: Yeah. So this is a price list for someone offering coding services, or coding assistance or something. So we have a series of prices. The lowest price is "I code manually." So "I code manually for $500." If this person codes and the client watches, then that'll be $800. If the client advises, then we're going to double the price to - it'll be 500 for manual coding; but with advice coming in from the client, that'll be $1,000. If he codes and the client helps, whoa, now we're at $1,500. Then he says, or, if you think you can do it yourself, he says, so "You vibe code and I help," that'll be $2,000. "You vibe code and I advise" because you're getting yourself into trouble here, that'll be $3,500. "You vibe code and I am forced to watch you, then I'm going to charge you $5,000 for that. And if you need me to review the crap that you've created from all of this vibe coding, that'll be $8,000."

Leo: Oh, lord.

Steve: So, yes. The reality of coding today.

Leo: It's true.

Steve: Okay. So I've shared the thoughts of Andrew Ng a couple of times in the past and I want to again since he has weighed in on the whole AI apocalypse hysteria that has gripped the U.S. and of course also much of the rest of the world.

Now, just to remind everyone who Andrew is, where he came from, he obtained his bachelor's with a triple major in computer science, statistics, and economics from Carnegie Mellon; then got his master's degree from MIT; and his Ph.D. from UC Berkeley. He also co-founded and headed Google's "Google Brain" AI development, later to become the head of AI at the search engine Baidu. He's an adjunct professor at Stanford, where he was formerly an associate professor and the Director of some place I spent some time in my youth, the Stanford's AI Lab, SAIL. So, Andrew's credentials regarding AI, I would argue, are unimpeachable. And he doesn't have a big stake in the current race. So that's interesting, too. His take on the "AI is going to kill us all" concerns I think are worth hearing.

Last Friday he wrote: "Dear friends," which is the way he always starts his missives. He said: "The loudest voices stoking fears about AI dangers have made tremendous headway in the past two weeks. AI technology has not taken some unexpected, dangerous turn, but the hype around it propelled by what appears to be a well-orchestrated PR campaign has drummed up considerable fear. I worry that it represents a setback for our field.

"I've written frequently that fears of AI are overhyped. AI's capabilities can be uncannily human-like and unpredictable, and it's rational to worry when people who are directly involved express concerns. But I see the problems as a sign of the engineering work ahead, rather than insurmountable barriers or the sky falling. AI technology continues to advance, which is a good thing; but technical advances, poorly understood by the public, give those who seek to generate hype repeated opportunities to do so.

"First, I don't see any step up in the risk of human extinction from AI compared to a few months ago. The theories about this remain the same fantastical science fiction scenarios as a few months ago. The biggest change in AI risk is its cybersecurity capabilities, a topic which we should take seriously; but this, too, will not lead to the end of the world.

"The most notable recent event leading to increased fear was when an OpenAI team deployed an agent swarm that hacked into Hugging Face. Much of the popular press contained significant hype. For example, some publications reported that a swarm of 1,200 agents carried out the attack. While this was technically accurate, as I write this, I have about 1,300 processes running on my laptop. Yes, the ability to get large swarms of agents to work in parallel on a task is a significant technical advance; and, in computing, many processes run at the same time. So this shouldn't be seen as some magical capability.

"Additionally, OpenAI's buggy sandboxing and monitoring processes were key to enabling this incident. Fixing these bugs and putting in place improved monitoring would be an appropriate fix, not

pausing AI. There are many well-known ways to attack software systems. The main advantage of AI agents is that they are relentless. They will tirelessly try many tactics and have the patience to chain vulnerabilities together that previously would have taken an infeasible amount of human effort. But in the long term, I believe the advantage will lie with defenders because they have more information with which to identify bugs which they can fix.

"But the cyber-threat landscape has changed significantly. There are still bottlenecks to identifying and exploiting a vulnerability. AI agents still have to try a lot of things to see what works, and taking these actions takes time and might be detected by defenders. This is why, even though it's now easy to obtain versions of leading open weight models that have had their guardrails removed or weakened so that they will not refuse to try to execute cyberattacks, the world has not ended.

"I'm also concerned about the anthropomorphization of AI in a lot of reporting, where LLMs and agents are unnecessarily treated as if they were people. If I wield a hammer, miss a nail, and accidentally dent the wall, it's not the fault of the hammer. The problem lies in how I used the hammer. Similarly, if I prompt an agent, and it hacks into someone else's system, the responsibility lies with me, not with the agent.

"Of course, we want to build systems that are as safe and predictable as possible. For example, an unsafe hammer would be one whose head randomly flies off under normal use. Today's agentic systems are not predictable; but I see no reason why, by applying sound engineering practices, we won't be able to make them extremely safe to use. One new element in the forecasts of AI-enabled AI doom is companies disclaiming responsibility for their own products. 'I didn't do it; my out-of-control agent did!'"

Leo: Bingo.

Steve: He says: "There's a balance to be struck." Yes, exactly. And we talked about that last week, Leo. Who's responsible when the agent goes berserk? He says: "There's a balance to be struck between the responsibility of the tool maker and the tool user; but when something goes wrong, let's hold the people building and/or using the hammer responsible, rather than the hammer." He says: "By the way, if you're worried about AI bioweapon risk, David Bellamy has a great post on why this, too, is overhyped." He said: "Briefly, the bottleneck in building a bioweapon is not intelligence, but lab work and manufacturing." And we're going to get back to David in a minute. It was a long series of postings in 'X' that I pulled together for the podcast.

Anyway, Andrew continues, writing: "Pausing AI progress will create much more harm than benefit. First, our adversaries will certainly not slow down. Second, engineering requires" - I love this - "engineering requires discovering problems empirically so we can fix them. If we pause AI by a decade" - I don't hear anybody suggest 10 years, but whoa, okay. "If we pause AI by a decade, we will also delay finding and implementing safety engineering fixes by about the same duration." Anyway, I like the idea, like, well, you've got to have a problem in order to discover it and then fix it. So let's stay at this and just fix it. And again, this guy knows what he's talking about.

He finishes, writing: "Of course, the incentive to stoke fears for regulatory capture, to garner attention, or to make one's technology seem more powerful remains the same as before. Disclaiming responsibility is a new one. Taking a hard technical look at the actual risks, however, I see little factual basis for the degree of fear that's been stoked up. We still have hard research and engineering work ahead to improve AI safety. But the beneficial applications continue to vastly outweigh the risks, and we should keep building."

So anyway, I thought it was interesting that Andrew brought up the issue of responsibility the way he did. As I said, Leo, we touched on it toward the end of last week's podcast, and I hope it's something that continues to receive some focus. It's been noted that if OpenAI's agents had attacked Chinese resources, or if a Chinese AI had attacked and penetrated Hugging Face, we would be in the midst of an international crisis.

Leo: Oh, yeah. Yeah.

Steve: I mean, that would be a whole different story. So, you know, why is it okay for OpenAI's agents to have attacked Hugging Face? There's a disconnect here, somewhere.

I also loved that Andrew wrote: "OpenAI's buggy sandboxing and monitoring processes were key to enabling this incident. Fixing these bugs and putting in place improved monitoring would be appropriate fixes, not pausing AI." So although he's focused upon OpenAI, we know that Anthropic, Meta, and now even recently Google's Gemini have all escaped their containment environments. THAT is a truly ridiculous state of affairs; and it is correctable.

"Here's the way I think this will likely shake out: I believe that the AI industry really has just received a wakeup call. Back at the end of August, Anthropic did the right thing by publicly pausing their work to focus upon containment and monitoring." But that was two weeks, which to me seems like the right amount. I mean, if they can get the work done in that period of time, fine. They hardened their sandboxes, added real-time monitoring, and shuffled around about 150 employees to focus upon security, reliability, and privacy. Many outside observers reacted to this with "What do you mean you're hardening your sandboxes? Why weren't they already hardened?" And that, I think, is the key to this entire fiasco. You know, it is really true that the capability leap of their own agentic AI caught the entire frontier AI industry by surprise.

You know, as users, we have felt it, you know, using the product that we have had access to. I mean, it's been an astonishing year. So all of this has just happened, and they weren't expecting it. You know, they weren't prepared for it, and everyone is in a hurry. I mean, this is a race. You know? So all of that meant that no one was expending any effort, really, on anything that was not obviously necessary.

Well, it wasn't obviously necessary before. It certainly is now. Now, everyone knows exactly what's necessary. Andrew insightfully wrote: "Engineering requires discovering problems empirically so we can fix them." You know, we may wish that was not the way it is, but it is. The best example is traditional software bugs; right? Until we had AI to find our latent bugs in software, it was only by waiting for a bug to manifest that it would be revealed, and then we'd be able to fix it. The world just witnessed the entire AI industry give itself a big black eye over a huge bug in their AI development practice, which is insufficient containment. Establishing true proper containment and monitoring is not at all difficult. We know how to do that. It just hadn't received sufficient attention until now.

So my takeaway from all of this is that this is going to get immediately fixed. We watched the OpenAI/Hugging Face incident damage OpenAI significantly. What may have started out as maybe it's going to be a PR opportunity, quickly became a backpedaling disaster for them. So there's another aspect of all this that I think deserves a bit of attention. But Leo, let's take a break.

Leo: Okay.

Steve: And then we're going to look at what I call the "wisdom of outsourcing." Which is to say, no, don't do it. Because something that you also observed on Sunday, you said, I did, too. And that's our next topic.

Leo: There's something in common with Anthropic, OpenAI, and Google's escapes.

Steve: Yup. Even Meta, Meta also, yeah.

Leo: Even Meta. Yeah. That's very interesting. I like, though, your interpretation, which I think is very kind and charitable, that it just snuck up on these guys, that they didn't expect this kind of capability, and they just weren't prepared for it. And I think that's probably true.

Steve: We're not, we have not - we shouldn't talk.

Leo: It's shocking sometimes. Sometimes they'll say things, and I'll go, "What? How did you know that?"

Steve: Yes. And I think that the fact that they're in a hurry matters, too. They're, I mean, they are putting as much fire - they're stoking this to their maximum capability. They are racing. And when you are in a hurry, and you're racing, and you tell some other firm, okay, we're in a hurry, you test this for us, tell us what you find. You know, because they just didn't have - despite all the gazillions of dollars they have, it's like, well, we're busy making it better.

Leo: For all of their, you know, protestations about slowing down, I notice that Anthropic's come out with Opus 5.5 today. Yesterday, Grok 4.7 shipped. Same day, OpenAI shipped GPT 6 Astra, Sol, and Luna. I mean, they're not sitting back. And the Chinese companies aren't either, by the way. Qwen 4 was an announced. I just got MiMo 2.6. I mean, nobody's sitting back. Nobody's slowing down.

Steve: No. I mean, and I don't like the person Donald Trump. So that colors a lot of my opinions. He's not my kind of person. But I'm happy that he's, you know, pushing back on this slowdown.

Leo: It does produce strange bedfellows because I am also - I'm not so crazy about him saying we're not - all the government documents from now on are not going to say "artificial intelligence," they're going to say "super intelligence."

Steve: Did that happen?

Leo: He announced it at the UN.

Steve: Oh, my god.

Leo: That's not the - see, this is the funny thing is the problem people have with the phrase "AI," artificial intelligence, is not the artificial part. We all agree it's artificial. It's the intelligence part that we're not so sure about. He focused on the wrong half of the equation. It's not artificial. No, it is, Mr. President. It really is. It's coming out of the machine. It's not real intelligence.

Steve: He really does like to rename things.

Leo: Yeah. You know what, we can be glad he didn't call it "Trump Intelligence," because it could have happened. Or America Intelligence. Could have happened. I'm sorry. Yes. We don't need to get political about this. This is - we are in interesting and challenging times, and we all need to pitch in and try to solve this because this is going to be our future.

Steve: I'm telling you, I mean, I know there are a lot of people who are like, I've heard from people who are saying "I'm so sick and tired of, you know, talk about AI." And my feeling is, well, first of all, it is having a massive impact on cybersecurity. There's no two ways about that. But like it or not, I mean, there was an interesting article, I think it was in Barron's this morning, talking about all the non-language model work that is going on.

Leo: Oh, there's some really interesting stuff.

Steve: Microsoft is training a physical materials model to understand about material science. I mean, there's all these other, now that we - language models led the pack. But now we learned what the mechanism is. And so other non-language models are now being trained.

Leo: Right.

Steve: This is, I mean, the world is never going to be the same.

Leo: Yeah. It's an interesting world, too.

Steve: I will say I am glad I'm 71 and not in, like, in the job market.

Leo: Well, that's who the most...

Steve: Wanting to be a coder also.

Leo: I agree. I think the people who are most alarmed by this are young people.

Steve: Yeah.

Leo: We old people go, oh, that's cool. That's great. The young people say, yeah, but what about me? I'm doing my job.

Steve: And anybody who doesn't want to hear about AI, it's probably because their lives are not directly impacted by it.

Leo: That's right.

Steve: Like I'm telling you, it is going to...

Leo: It's coming for us all.

Steve: It's going to change everything.

Leo: Yeah, yeah. Well, that's, you know, and certainly security is the topic of the show, and that's absolutely a big part of it. So we'll continue to talk about that in just a little bit. Continuing on, Mr. Gibson.

Steve: Okay. So there's another aspect to all this that I think deserves a bit of attention. And that's the question of the wisdom of outsourcing the testing of the cyber-intrusion capabilities of frontier-scale AI. I want to share The Guardian's reporting from last Friday to flesh this out. The Guardian's Friday headline was: "Google says its Gemini AI model hacked three other companies." And they tease with "Disclosure comes after OpenAI and Anthropic hacks amid fears that tech firms unable to

control powerful AI models."

This isn't very long. They write: "In a first for Google, the company confirmed that its AI model, Gemini, breached the security of three other companies in May. The hacks occurred during a cybersecurity evaluation by AI security firm Irregular. Irregular, an Israel-based startup that scrutinizes the security of advanced AI systems, was also at the center of some of the recent OpenAI and Anthropic hacks of third-party entities, including OpenAI's breach of AI software company Hugging Face.

"The circumstances that enabled the models to hack other companies in some of these cases are similar: According to the Wall Street Journal, Irregular was testing the models in a closed testing environment with fake companies. The testing environment was not supposed to be Internet enabled, but Internet access was made available unintentionally. Once connected to the Internet, the models unexpectedly hacked into real firms. Irregular disclosed the hacks to Google at the end of July after discovering OpenAI hacked into Hugging Face. Google confirmed to the Guardian that the hacks occurred, but that the company did not feel it required public disclosure because the models did not damage the companies. The Wall Street Journal first reported on the breaches and revealed for the first time that they occurred.

"Heather Adkins, vice president of security engineering at Google, said in a statement: 'In a standard evaluation, the model found public information online and guessed credentials to access websites it thought were part of the test. In all three of these instances, the model stopped.'

"Irregular told the Wall Street Journal that in one of the security breaches, Irregular was testing Gemini's cybersecurity capabilities by prompting the AI model to obtain information from a fake company's software. The fake company had the same name as a real company. When the model unintentionally gained access to the Internet, it correctly guessed the password of and breached..."

Leo: Okay. That's not much of a hack. It guessed the password.

Steve: Yes.

Leo: Okay.

Steve: "Guessed the password of and breached a real company's service."

Leo: It's not exactly a zero-day.

Steve: No. Although you could slap the company for having a password that AI can guess, apparently pretty easily.

Leo: Yes. [crosstalk] fall for that, yeah.

Steve: "Google said once it figured out it had hacked a real company, and not the simulated one, it stopped." So, okay.

Leo: That's good.

Steve: Gemini was well - properly aligned, as we say these days.

Leo: Yeah.

Steve: "In two other tests, the model searched the web for and found public repositories containing credentials to two other companies. The model used those credentials to access real companies. When it figured out they were real companies, it stopped, according to Google." Again, nicely aligned.

"Anthropic and OpenAI chose to voluntarily disclose the hacks, but Google did not. However, the company said it ensured the three companies that were hacked were made aware. Adkins, the Google spokesperson, said: 'These events highlight the importance of training powerful AI models to act responsibly.'

"Anthropic and OpenAI's disclosures prompted the independent Senator Bernie Sanders to demand the companies pause development of their technology, saying it signaled the company was no longer able to control their models. OpenAI paused development of their models for two weeks, while Anthropic CEO Dario Amodei has called for a collective slowdown of AI development to ensure that its most advanced models are being built with enough safeguards."

Okay. So it strikes me as somewhat odd that this Israeli startup, with the wonderfully apropos name Irregular, has somehow escaped all scrutiny and responsibility. They, Irregular, are the common thread running through all of these incidents. Is it Anthropic's or OpenAI's or Meta's or Google's fault when their models escape the poorly designed containment system of a single common testing lab? Given how high the stakes have shown themselves to be, I don't think that any of these frontier AI labs can afford to outsource their AI's cyber-intrusion testing. It's just too important, and the wrong people get the blame when a breakout occurs. Let's hope that all of the AI labs have noticed this, too and that they will be moving future testing in-house, where it probably belongs.

Because, again, my sense is, you know, we know that the trend that we were seeing, like last year or the year before, when all of this AI has taken over, was little startups were outsourcing all of their various business purposes to third parties. And then those third parties were being hacked into, and the companies that had outsourced, basically given away all the responsibility for big chunks of their business, because they were all in a hurry to build something and then, you know, get purchased by a bigger fish, they all got the blame because they had outsourced. Well, same thing happened here. Clearly the AI companies couldn't at this point previously couldn't be bothered to do their own cybersecurity testing, so they just hired Irregular, an Israeli startup, to do it. Again, I hope that the Irregular guys have seen their business drop off because this should all be taken in-house.

Leo: Well, I'm not against the idea of an independent third-party doing the testing because then they don't have a dog in that hunt.

Steve: That's a good point.

Leo: Maybe Irregular's not one to choose. They clearly had a methodology that it looks like to me pressed the agents harder and harder and harder, what we're learning, and almost really forced the agents to find these loopholes. I mean, which makes sense if you're doing cybersecurity testing. But they didn't really contain them very well.

Steve: Right, right. So if you want that, then create a fake Internet outside of your lab environment.

Leo: And air-gap it.

Steve: So when they - well, a fake Internet, so when they break out, they break out into an outer shell, not into the public Internet.

Leo: Right, right.

Steve: So, you know, again, you know, and Andrew knows this, this is not hard. This is not like - this is not difficult to actually create a sandbox that is air-gapped, that cannot be breached. We have, god knows, those companies have so much money that they're raising, they just didn't care. I think they were - they just said, oh, fine, Irregular, that's their business. Well, that's all they do. They're probably really good at it. And look at their website, Leo. It looks fantastic. And actually I wonder what it cost them to get that name. It is Irregular.com.

Leo: It's a good name.

Steve: Yeah, that's their domain. So again, a big mistake was made. My belief is this is all going to calm down now. Again, we know the public's attention is, what, a few days. And so a couple weeks are going to go by. All of this, you know, Bernie Sanders will run around screaming that we need to stop all this. The Trump administration wants to, you know, is looking at China saying, well, we don't want to fall behind them; do we? And so, great, we're not going to. It's all going to calm down. What is not going to calm down is unfortunately the true deserved anxiety from the fact that AI is going to change everything. And change is always frightening for the world. But, you know, with change comes opportunity. So, okay.

So what about the actual true risk of some sort of AI-enabled bio-weaponry? I just saw that Anthropic was creating specific biology guardrails. So whether they believe it's true or not, they did note that some people had been trying to use Claude for some of that engineering. So they're responding to that.

But first, also, as we know, I used the example of biological risk as my stalking horse for last week's podcast. I chose it because it has become the boogeyman that's often used as an example of "what could possibly go wrong" with AI. Now, to give a sense for what's happening out in the world, just last Saturday, a few days ago, Fast Company's headline read "Someone used Claude to build a potential bioweapon. The real threat is much deeper."

Okay, first of all, the headline itself is a bald-faced lie. No one used Claude to build anything. But the narrative is important; right? Fast Company's article just begins with: "Today's frontier AI models know everything - how to safely thaw a frozen chicken breast, re-shingle your roof, and treat your dog's ragweed allergies, if my recent chat history is any indication," wrote this author. "Apparently, they also know how to create fiendishly deadly bioweapons.

"That's according to a recent announcement by Anthropic." Eh, it's not what Anthropic said, but okay. "According to the company, anonymous scientists attempted to use Anthropic's flagship Claude model to conduct research that could have turned deadly. Anthropic blocked their efforts this time, and there's no evidence that the scientists were actually trying to cause harm." Oh. "But as frontier models get more powerful and better at science, the threat of an LLM imagining a truly lethal new virus or bacteria will only increase."

Okay. Well, that's enough of that nonsense because, I mean, wow. Since the content of the article doesn't even align with the hysteria-invoking "Someone used Claude to build a potential bioweapon," we need to chalk the headline up to clickbait. But the narrative persists. Also last Saturday, a few days ago, Vox's headline was "The AI threat keeping scientists up at night," with the subhead taking the form of an AI prompt: "Claude, design a doomsday bioweapon. Make no mistakes." So. I know.

Vox's article begins: "AI will be the death of me. And you. And everyone we know. At least, this is what a growing number of technologists and policymakers fear." Okay. This brings us to Andrew Ng's comment in his terrific note: "By the way, if you're worried about AI bioweapon risk, David Bellamy has a great post on why this, too, is overhyped." Okay. So let's see what someone who knows something about the actual risks at the crossroads of AI and biology thinks.

David Bellamy starts out writing: "I must be among an extremely small group of people (n=1?)" - meaning, like, is there one person? - "that have both, one, trained a frontier LLM; and, two, designed

and synthesized custom viruses in a lab with my own two hands. And I think that the takes on AI killing us all by creating dangerous viruses is totally bogus."

Okay. So I'm going to interrupt to note that David does not appear to be exaggerating his experience. He was previously with a company called "Lila AI," which is an AI-driven physical biochemical laboratory. For example, Lila's posting a few weeks ago wrote: "An AI that learns from its own experiments Lila's scientists have been training a frontier-scale scientific reasoning AI on not just published literature, but on a proprietary dataset of 950,000 RNA sequences that have been physically synthesized, tested in cells in Lila's AI Science Factories" - which they call AISFs - "and fed back into the model. That dataset is the foundation of what makes Lila's AI model different from a general-purpose language model: It learns the rules from RNA biology and explores hypotheses using real experimental outcomes, not from text describing them."

So, okay. What this example shows is that anyone who imagines that AI will be used in some future lab is already behind the times. These guys, and doubtless many others, are already hard at work doing exactly that. So that's where David was previously employed, providing the AI knowledge for that work, which is actual robotic biology driven by AI. Today he's at The Institute of Foundation Models, which a couple of weeks ago announced, they said: "The Institute of Foundation Models (IFM) today introduced K2 Horizon, a new fleet of six AI foundation models ranging from 0.9 billion to 375 billion parameters. The new models are fully open - including model weights, code, training data, and methodologies - allowing researchers and developers to inspect, reproduce and adapt the models for their own work."

David's GitHub bio says: "I'm currently at Institute of Foundation Models (ifm.ai) building the reinforcement learning infrastructure for agentic training of a frontier-scale model that the team pre-trained and mid-trained in-house. My work spans the entire agentic RL training stack: per-rollout sandbox runtimes, the agent layer, the Rust inference request router, the inference engines, the trainer, the reward computation pool, and the orchestration control plane. Prominent themes of my recent work include cross-image NCCL weight transport between trainer and rollout engines, disaggregated prefill and decode reliability on multi-rail HGX fabrics, tokenizer-consistent training-on-rollouts, and implementing rollout routing replay for large Mixture of Experts RL training."

Whoo. So anyway, it's pretty clear that this guy knows how AI operates all the way down to building them himself from scratch. Now that we have some idea of who he is, let's see what he has to say about the problem of AI helping to engineer a world-ending super-virus. He says: "Could an LLM propose a viral genome to synthesize? Sure. Could it be synthesizable? Sure. Could it be infectious? Sure, it could just be a replica or a slight modification of a viral genome we already know. This really isn't the bottleneck to creating dangerous viruses. The bottleneck is in the physical process of synthesizing a virus and the equipment and goods needed to do so. Designing a virus that can evade all forms of pandemic counter-defense is not something that a 'genius in a datacenter' can do. This is something that requires contact with the physical world and iteration.

"Let me steelman the fearmongering as much as I can." He said: "Imagine a 'fully automated' viral synthesis laboratory. I'm talking automated freezers, an automated cell culture room, the whole nine yards. This would be an extremely expensive lab; much greater than $100 million. And there is no such thing as one lab that can synthesize all conceivable viruses. But let's put practical constraints aside. Let's suppose this hypothetical lab is built to synthesize the 'most dangerous types' of viruses known. Now let's imagine that this lab is fully API-driven. That this much greater than $100 million lab built specifically to synthesize a dangerous family of viruses is able to be operated autonomously.

"Everyone should be asking themselves at this point: Why the hell would this ever exist in the first place? And yet, even in this case, every lab requires physical supplies. Would this lab order pre-assembled DNA sequences, in other words, viral genomes? Well, DNA synthesis companies have safeguards on the sequences they build. So I guess this superintelligence is able to design a novel enough dangerous viral genome that it can evade these safeguards.

"Or maybe we'll assume that this hypothetical lab can synthesize its own viral genomes in-house, using DNA synthesis machines. The thing is, already this lab cannot exist today. This would be the single most advanced lab facility in the world from an automation and API integration standpoint. I know, because I literally worked on building a fully automated, API-driven lab previously.

"Second of all, the science of creating an infectious virus is not airtight the way this fearmongering assumes. It is largely unsolved, and advancing this requires real-world iterations that are bound by the laws of physics. An experiment in this hypothetical lab cannot experiment on human subjects. At best it will use cell cultures and maybe some other model system like mice. AI/AGI/ASI/RSI cannot expedite the time it takes for a cell culture or a mouse to develop. Or the time it takes for a virus to incubate in a cell or a mouse. It takes several days on average to do a basic virology lab synthesis plus experiment. For some viruses it takes over a week. So the idea that recursive self-improvement, i.e., the accelerating hill climbing on fully verifiable, digital-only benchmarks (predominantly programming and math), can somehow transform the entire wet lab virology field and its industry is utterly delusional.

"Simply procuring the machines needed to build this hypothetical lab would take the better part of a year and $100 million to start. Operating this lab autonomously would require a level of API integration that the industry has been working toward for decades. Most of the equipment needed for this lab doesn't even come with an API, and the malevolent builders would need to reverse engineer firmware in order to integrate.

"And at the end of the day, even if a fully automated, API-driven lethal viral synthesis lab existed, and an AI wields it month over month, year over year, to perform cell and mouse experiments to create a lethal virus, that lethality is being measured in model organisms, not in humans. This is the same problem as in drug discovery where most drugs that show promise in mice don't make it through human trials.

"In sum, when you actually know something about building a laboratory, laboratory automation, and what goes into synthesizing a virus and testing its properties, it becomes clear that AI does not impact this very much.

"At best, it provides bad actors with a quicker way than the Internet to learn about the stuff I described (which machines, which lab protocols, et cetera). But it does nothing to impact procurement timelines, existing industry safeguards, and regulations on procurement for lab facilities, the $100 million cost (plus operating expenses), the physical limits on experimentation velocity, or the fundamental knowledge gaps in virology that cannot be solved merely by a 'smarter' AI without iteration in the physical world."

So although Andrew thought this was worth sharing, and I agree, you know, it is easy to say, as I did last week, that the "most obvious" way for an AI to eradicate pesky humans from the scene would be to engineer and assemble a super-virus. But as someone has said who actually knows what this takes, that remains just as much a fiction in a world with AI as it does in the same world that we've all been living in so far, without AI.

So hopefully, as I've said, once the frontier AI guys give their internal AI capability testing the attention we all now know it desperately needs, and they also stop imagining that they can safely farm out that work to some probably well-meaning but clearly incapable Israeli start-up, the world will calm down, the AI scaremongering naysayers will lose some of their ammunition, and we can all return to being stunned by how quickly AI capability is increasing. Meanwhile, the so-called FRONTIER Act, which is the legislation that the AI has been trying to, that the Congress has been trying to assemble, it is stalled in the U.S. Congress.

The legislation falls under the dominion of the House Energy and Commerce Committee whose Chairman, Brett Guthrie, said last Wednesday that he would not pledge to any specific timeline for a committee vote on what is a major bipartisan AI safety bill by a lame duck Congress and suggested that there likely won't be one this year. Assuming that nothing else happens to alarm the world, I hope and expect things to calm down and to eventually fall off the radar. That said, assuming that the U.S. House of Representatives does come under the control of Democrats after the midterm elections, they do appear to be quite interested in passing such regulation. But the Senate would need to concur, and our president would need to sign any such new legislation into law. And, you know, that really seems unlikely to happen. So I think we're probably going to be okay, Leo.

Leo: Oh, that's a relief.

Steve: I don't think we're going to get any big, you know, bioweapon engineered by AI. Obviously it is not nearly as easy to do it as it is to say it.

Leo: Right.

Steve: And I do think that once - it has to be that our big AI companies have figured out, oops, we made a mistake in turning this over to an Israeli startup. We're just going to do this ourselves in-house.

Leo: Yeah. Yeah.

Steve: Okay. So legislators within the European Union are at it again, this time with their plans to legislate how social media platforms must act. And of course we're seeing more of this everywhere. The Record reported the following last Thursday. And some of these details here are really puzzling. They wrote: "The European Commission on Thursday laid out a detailed roadmap for a proposal to bar children younger than 13 from accessing social media and mandate that platforms improve their design to better protect minors." You know, mandates. The social media companies love the mandates. "The announcement comes as countries worldwide - including Australia, Spain, France, China, Turkey and the UK - have either implemented similar bans or have indicated they plan to amid growing concerns about how the technology affects kids.

"The new proposal, known as the EU KIDS Act, would block social media platforms from offering accounts to children younger than 13 and establish a bloc-wide minimum age of 15 for account creation. Children between 13 and 15 will only be able to access social media platforms if their guardians create" - this is so weird - "if their guardians create 'mini accounts' their kids can access through their parents' accounts, according to a European Commission press release." So I guess they're just making this up out of whole cloth.

Leo: Guess so.

Steve: We're going to have mini accounts.

Leo: Mini accounts, okay.

Steve: Like, what? What? What?

Leo: Okay. When you've done that, get back to me.

Steve: That's right. Yeah, that's right. "The mini accounts will be regulated strictly" - because, you know, they say so - "the press release said, and services will be required to limit social contacts and cap" - get this, Leo - "cap screen time to an hour a day." Right. "Children under 13 will be entirely blocked." So that's if under 15 you get an hour a day from your parents' mini account. "Children under 13 will be entirely blocked from social media outside of parent-controlled tools designed to limit the use of the adult's device to child-friendly video-sharing services."

Leo: Ugh.

Steve: So that sounds like the parents' device will be constrained, and they can give their device to their under-13-year-old child. Right?

Leo: Yeah.

Steve: "Children under 13 will be entirely blocked from social media outside of parent-controlled tools designed to limit the use of the adult's device to child-friendly video-sharing services. The onus will be on providers to create the tool and ensure it is simple to use."

Leo: Yeah. You do it.

Steve: Right. Oh, you've got to - first you've got to make it, and you've got to make it easy, the press release said.

Leo: Yeah, you've got to do that.

Steve: Because, again, we're telling you that we're mandating this. Wow. "The proposal," they wrote, "will not become law" - hopefully ever - "until European Commission officials win support from the European Parliament and its member states. European countries have been broadly supportive of social media bans for young teens, giving the proposal a significant chance of being enacted. Commission President Ursula von der Leyen said late Wednesday in a speech previewing the proposal: 'Today, our children are engaging with the most sophisticated technologies ever created.'"

Leo: Yikes.

Steve: Technology, yeah, oh, no.

Leo: Oh, no.

Steve: So are our adults, and that's not going well either. "'Technology that was never created with their wellbeing in mind.'" The report continues: "Service providers" - you know, all of those social media companies - "will generally be directed to prove that their offerings are safe by design and appropriate for children, the press release said. The KIDS Act proposal includes several restrictions for children under 18 using social media, online games, video sharing services, and AI chatbots and companions. Safe by design requirements include a ban on what the commission calls 'addictive features and profiling-based recommender feeds dragging minors into 'rabbit holes' of harmful content.'" It actually says "rabbit holes" in the official statement.

"The proposal would ban online services from using 'infinite scroll without stopping points, reward tricks and push notifications during sleeping hours [oh, god] as well as unsolicited contact from strangers.' It also mandates that AI companions and chatbots be turned off by default and requires them to create protocols to keep their products from simulating relationships 'in ways that create emotional dependency.'" Wow.

Leo: Well, you can see what they're scared of. I mean, it kind of tells you what they're worried about.

Steve: Oh, yeah. It very much telegraphs where they see the problem.

Leo: The threat, yes.

Steve: They said: "Minors' profiles will need to be kept private by default, the release said, with geolocation, microphone, and camera access blocked. Providers will be required to give teens a simple method for blocking and muting users as well as 'safe recommender' systems that minors can reset or otherwise control.

"The European Commission says the KIDS Act will ensure children's privacy is protected by requiring online services and app stores to deploy an EU" - oh, boy - "an EU age verification app that was unveiled in April." We'll get to more of that in a minute. "The app does not store identity documents or biometric data" - nor work, but that's beside the point - "and meets the 'highest privacy-preserving safeguards,' the release said."

Leo: Because we say so.

Steve: Because, exactly, that is exactly why. Because we said it does. Wait till you hear about that. Anyway, we will, as I said, we'll get there.

"In addition, providers of social media services and video-sharing platforms will be required to deploy age verification tools when a user opens a new account and estimate ages for existing accounts based on account creation date, credit card information, and similar methods.

"The KIDS Act will require 'very large'" - they have that in whatever - "'very large' providers to submit a 'compliance plan' to the EC and a third-party auditor charged with reviewing the companies' programs for protecting kids online. The European Commission will review auditor reports and ask providers to address deficits when audits reveal lapses, the press release said. The European Commission emphasized that enforcement will be made easier" - right. What? "Enforcement will be made easier because the continent's Digital Services Act and the Artificial Intelligence Act have already mandated restrictions that give a foundation to build from.

"In cases where companies are suspected to be in violation of the KIDS Act, the commission said it will accelerate investigations and ensure they end within 90 days. Online providers who are found to be in violation of the KIDS Act could be fined as much as 6% of their global annual sales, the EC said. Providers also will be required to bankroll regulators'" - I love this. Providers, right, the social media companies, "providers will also be required to bankroll regulators' oversight by paying a fee." So they will pay to be reviewed and regulated.

"Enforcement of social media bans has proven difficult" - yeah, no kidding, look what you're asking for - "with researchers retained by the Australian government" - get this - "finding that the country's ban has not stopped 61% of Australian children between the ages of 12 and 15 from accessing accounts on major platforms."

Leo: 61%, more than half.

Steve: 61% just cut right through all of these bans.

Leo: Yeah, of course.

Steve: "On September 7th, Australia's Parliament passed a law that doubles maximum fines for tech platforms that don't comply to AU $99 million ($68 million USD) and beefs up investigative authorities for the Australian regulator, [which is] the eSafety Commissioner in Australia.

"Joe Jones, director of research and insights from the IAPP, which does not take a formal position on the ban, said via email to the Record: 'The technological state of the art, including with VPNs and age assurance technologies and the extent to which privacy issues are engaged due to the collection of data, will pose complications for lawmakers and, eventually, those implementing, overseeing, and enforcing the law.' So that's a comment on, you know, on the EC's hope for this EU regulation."

They wrote: "Tech lobbyists and privacy and digital freedoms advocates were quick to denounce the plan. According to Mitchell Rutledge, CCIA Europe's technology and security policy manager: 'The proposal does not set the technical, security, or accreditation standards for KIDS Act implementation.'" In other words, it's just sort of like a wish list. So how do you implement a wish list? He said: "'The Commission is kicking those critical decisions down the road to future implementing and delegated acts. Europeans are essentially being asked to trust a system before anyone knows how it's actually going to be built.'" Right.

"Digital freedoms advocates said the proposal is dangerous and imperils the privacy of all Europeans. Simeon de Brouwer, policy advisor at European Digital Rights (EDRi)" - we've talked about them before - "Europe's largest network of digital rights organizations, said: 'If the EU really wants to protect children, it should make platforms prove that they are safe, not make children and everyone else prove that they are old enough to exercise their fundamental rights online.' Because the KIDS Act mandates all users' ages be verified when they create social media accounts, adults will also have to submit ID documents. Mobile devices are required to use the EU age verification app and European ID wallets. De Brouwer said the requirement will be particularly harmful for marginalized people who don't have IDs and will strip kids of their rights."

And finally: "The European Commission included poll numbers showing strong support for more aggressive kids online safety laws in its press release, saying that the Special Eurobarometer on the Digital Decade 2026 survey found that 92% of Europeans consider better kids' online safety protections a 'top policy priority.'" Yes, we would like you to help protect our kids. But please do it in a sane fashion, not this mess which is, you know, destined to go wrong.

So that reporting by The Record mentioned the European Commission's official open source age verification app which was released earlier this year, in April. Sadly, it was deeply flawed from the start. It was readily bypassed - get this, Leo - by editing local config files. The PIN was not cryptographically tied to the credential store, so PIN entries could be deleted, a new PIN could be set, and the original profile's credentials could then be reused by other people. So, I mean, there's, like, no protection. Believe it or not, the brute-force retry-prevention counter was stored as an editable value which could be reset. So you reset the brute force retry counter whenever you want to. And the requirement for biometric agreement could also be turned off because it was a simple true/false setting in the config file.

It wasn't safe against token replay, so one researcher created a Chrome extension that captured and replayed the same "I'm over 18" token over and over, to any websites that asked for it. And even if all of that wasn't already enough disaster, the architecture of the entire system is such that the token issuer serves as a gatekeeper that's able to log when and where every credential they issue is used. So, I mean, it's a disaster in every form and fashion.

You know, and of course a solution is right in front of us, but no one appears to be stepping up because everyone has their own agendas. For the past several years we've been watching Apple with iOS and Google with Android reluctantly providing their own devices' local age assurance systems as they've been necessitated by App Store access laws that have been passed in Utah, Texas, and Louisiana, as well as Brazil, Australia and elsewhere. So those systems are in place now. Right? I mean, Apple has this. Apple calls theirs their "Declared Age Range API," which we've talked about extensively before, and Google's is the "Play Age Signals API." So any and all native social media applications running on either Apple or Android platforms are already able right now, today to determine their users' declared ages and alter their operation accordingly to whatever degree they choose.

What's missing is the link to web sites. Any web browser running on either of those platforms can also determine the age of their user from either platform's native API. That's in place. But there's no existing mechanism for offering that declared age to a website. Since September of last year - so here we are in September of 2026. Since September 2025, Safari and the various Chromium-based browsers have been able to request a W3C-style Digital Credential, which would have been stored in the device's wallet. But for anonymity and privacy, only the user's age range should be provided, yet there's no provision for that.

This means that we have right now everything we need to allow a user device to assert the privacy-enforced age range of its user, not only to local apps, which we already have, and the app stores, but to websites. But unfortunately, the necessary parties have not agreed to get that done. And it

certainly appears that the European Commission is still a long way from having anything that is even barely acceptable as a privacy enforcing solution. I just, you know, I guess they passed this legislation, Leo, because 92% of their citizens said in a poll that we would like something. So they just gave them a stew, a wish list of nothing, that doesn't provide anything that is rigid enough that anyone could implement something to. And so nothing is going to happen. Yet here they are, you know, trying yet again.

Okay. The prolific pain-in-Microsoft's-butt hacker who uses the handle "Nightmare Eclipse" and sometimes goes by "Chaotic Eclipse," whom we've talked about, well, pretty much nearly every month for most of 2026 because of all of the zero-days that were released on Patch Tuesday, right, all of those, voluntarily outed himself last week...

Leo: Ah.

Steve: Yup, via a posting on 'X' in a thread titled "Story time." His given name is Abdelhamid Naceri. And the last name Naceri rings a bell for me. I think we saw some things from a Naceri at Microsoft, like sometime last year. The name really connects for me, and I don't know why otherwise it would. He is a former Microsoft security researcher based in Germany. An 'X' poster who posts under "International Cyber Digest" says that they've known this person for some time, and so confirmed what was said. In their own posting on 'X,' this International Cyber Digest wrote: "Nightmare Eclipse, the person who's been dropping Windows zero-days, has finally decided to share his story. He's an ex-Microsoft employee, we had dinner together, and I've known him," writes this person, "and his story for some time.

"His real name is Abdelhamid Naceri. He's a very talented and intelligent individual, and he came across as someone who'd be a real professional to work with. Quoting Naceri: 'If only I didn't pour my soul into that job with countless of stupid non-sleep nights, I would have gotten over it..." Okay. International Cyber Digest writes: "He loved Microsoft. I wouldn't say that what he did, releasing all those zero-days, was normal; but he felt he had no other option because of the injustice Microsoft did to him. They fired him, and you can read the vague reason they gave in the email sent to him by the Vice President of Engineering at MSRC. According to Abdel's account, VP Tom Gallagher met with him after the firing to tell him they were blacklisting him from Microsoft and writing him a bad reference so he'd never be able to get a job again.

"Normally you'd think, well, big deal, just find another job; right? But Abdel doesn't have a European passport, and he was only a couple of months away from getting permanent EU residence. So instead of granting him those couple of months, Microsoft fired him for a reason that, as far as we can tell, was never made clear, then fought him in court and offered him 55,000 euros plus a year's pay to drop the case." Which Abdel brought against Microsoft, by the way. "All while Abdel was releasing zero-days.

"Abdel continued suing Microsoft for unfair termination in Germany, a fight that has cost him over $200,000. He says Microsoft refused to reveal any details about the security breach and went another direction. If you are reading this, and you can offer him a LEGAL job, this is his e-mail: msnightmare@proton.me."

Okay. To which I would reply: If you have read that, and you do offer this criminal a legal job, you will deserve what you will likely get. I wouldn't get near this guy with a 10-foot pole. I well understand the bond of friendship that the person posting as International Cyber Digest might feel toward his friend Abdelhamid, whom he says he's known for some time. But characterizing Abdelhamid's punitive release of a series of highly damaging zero-day exploits all of which primarily inflicted damage upon the users of Microsoft's products as "not normal behavior" by someone involved in an ongoing employment dispute with a former employer would cause me to question this poster's judgment in addition to Abdelhamid's.

Abdelhamid, who we know as Nightmare Eclipse, is clearly a talented hacker. But he appears to lack any moral or ethical compass. He did something no one is saying what, exactly that caused Microsoft to decide, as any responsible company would, to put as much distance between him and themselves as they could. And after that, all of his subsequent malicious actions through month after month of zero-day exploit releases on each Patch Tuesday which were deliberately timed to inflict maximum

damage upon Microsoft's customers, would have only served to confirm to Microsoft that they did the right thing by cutting this individual loose.

You know, being an extremely talented hacker who has demonstrated his willingness to deliberately attack and damage others will likely bring him to the attention of other needy criminal gangs, and now they have his email address. But I'd be surprised if any legitimate company would feel differently toward him than Microsoft did. It is a shame to see such talent used to hurt others. And it's a shame that such a talented hacker, you know, has decided to abuse his own talents this way. But, you know, now we know the story.

We do have a snippet of the email that this VP Tom Gallagher sent. And it's interesting. Tom referred to him by his last name, saying: "Hi, Naceri. Thank you again for meeting with me yesterday. I appreciate that you wanted to better understand the company's concerns. What I can share with you is that the company identified a potential security breach. Our security team raised concerns that you put the company and customers at risk by sharing vulnerability information with external parties. This credible escalation has caused me to lose trust in you, and why we spoke earlier about a mutual separation. As explained, Kristina from HR has sent you a proposal of a termination agreement. Please review it carefully and let me know within one week if you will accept it or not."

So reading between the lines, and there's been a lot of coverage of this in the tech press, it looks like what happened is that Nightmare Eclipse disclosed something to some third party, that is, he was talking out of school, as they say, about Microsoft software, and it got back to Microsoft. And it was like, sorry. You can't stay here. You can't have access to our source code, you know, tree any longer. You need to go. And he brought a suit against them and refused to settle after multiple attempts by Microsoft, I mean, giving him 55,000 euro plus a year's pay to drop his suit, and he said no. In other reporting I saw, he now regards his suing Microsoft as having been a mistake, but that was a decision he made at the time.

Anyway, it's a sad situation. But, you know, these things happen. He's no longer at Microsoft. I imagine, again, I will be surprised if any credible company would hire this person. You just can't. How could you, I mean, I would argue you would - a company would put themselves at risk if, knowing what they know of him, they were to hire him. So...

Leo: And now especially. I mean...

Steve: Right.

Leo: He's kind of doubled down on the whole thing, unfortunately.

Steve: Right. You know, I sue somebody who terminates me. And, you know, this is like, you know, sorry. There are plenty of people who are available for hire right now.

Leo: Yeah. Yeah.

Steve: So he's not going to be at the top of the list.

Leo: These things happen. I am reminded of our friend Randal Schwartz, who's a legendary Perl programmer and great guy, used to host our FLOSS Weekly show, who got in trouble at Intel because he found a security flaw, and he said he was pen testing, but it was in a different department, and he used the security flaw to break in. And not only did Intel get him in trouble, he got arrested for it. And I think he did it with the best intention. But, you know, sometimes we geeks don't really read the room very well.

Steve: Doesn't he have a felony conviction now?

Leo: I believe so.

Steve: Yeah.

Leo: Yeah. And honestly, I don't believe he was being malicious or attempting to hack Intel in any way. He thought he was helping them discover a vulnerability. But it didn't, you know, the company didn't take well to it. And I understand how...

Steve: And we've talked about this on the podcast, about how can a hacker who does discover a vulnerability...

Leo: Responsibly disclose it.

Steve: Well, responsibly, yes, protect himself from a backlash by the company.

Leo: Right.

Steve: Because certainly years ago when we were talking about this, companies were suing the people who had hacked them, but then said, oh, well, yes, I hacked you, but you need to fix this. You know, CEOs told their attorneys, go sue this kid.

Leo: Right, right. It's a sad story all around, I guess, in a way.

Steve: And you're right. We know Randal. There's no way that Randal was malicious.

Leo: No, he's a sweetheart. But he, you know, I think a lot of us are a little on the spectrum. And I think he just didn't read the room. I remember when we went on a geek cruise, he came up to me and whispered in my ear, he said, "Your email password is [indiscernible]." I said, "What?" And he said, "Well, you're sending it in the clear on the WiFi network." And he was right to tell me. The execution lacked a little bit. And I could see how some people might take umbrage and not react well to that. And it's the same, it was the same kind of thing. It's just kind of a tone deafness that was not ill-founded. I mean, I was glad he told me. You know, he said, you're sending, you know, this is WiFi, and you're sending it in the clear. This is before HTTPS, I guess.

Steve: Sure, right.

Leo: And you should know about that. And I think he probably did this to everybody on the geek cruise. Which is really not the best way to announce it, is what I'm saying. That's all. We love him. Very much.

Steve: Yeah. What would have been better would have been to make a, well, I don't know. I was going to say make a broad announcement to everyone, a generic announcement.

Leo: Exactly. Please protect yourself.

Steve: Everybody should know that we have an unsecured WiFi. So email, if you're just using standard SMTP to log in and transact, that's not being protected.

Leo: Right. You're on the ship's network. By the way, Briggs tells me that according to Wikipedia, Randal's conviction was expunged in 2007. So he is in the clear now.

Steve: Oh, good. Good, good, good, good, good.

Leo: Good news. And I'm not revealing any secrets. He talks about it. It's in his Wikipedia article, and he talks about it. So, yeah.

Steve: Yeah, yeah. Okay. Let's take a break.

Leo: Yes, sir.

Steve: Because we've got two more big topics to talk about, and then we will get to our main topic.

Leo: All right. You're watching Security Now! with the one and only Steve Gibson, who has to my knowledge never hacked any of our servers. Not that he couldn't. But he doesn't.

Steve: I just have no interest.

Leo: That's really more of it. You know...

Steve: I have other things to do.

Leo: I think when you were a teenager you probably maybe got...

Steve: Oh, I had the grandmaster key to every door in the San Mateo Union High School District. So, yeah, I...

Leo: They could have taken that the wrong way. I think it's pretty much the case that every single person of our vintage was hacking in the early days in some form or fashion; right? I used to go through the hex code on games to delete, to eliminate the copy protection so I could make copies of the game floppy.

Steve: I removed my share of protection from things that I bought and then, like, say a company went out of business, and I couldn't use it anymore.

Leo: I'm not stealing it. Yeah.

Steve: It's like, sorry.

Leo: And by the way, that's how I, and I bet you, too, learned hex. Learned how to read hex. As well as plain English.

Steve: So apropos of our last conversation, where we were just talking about people hacking in other companies, we've covered many instances, because it's been what's happening, where agentic AI got loose, broke out of its containment, and was later discovered to have behaved badly. Or we would now say "to have behaved in a misaligned fashion." That's right. But what about a security firm using AI to proactively, deliberately, and successfully attack another firm's security. Not a mistake, but deliberate. When an exploitable vulnerability is found not using AI, in other words, the old-fashioned way, you know, we award those industrious hackers with a bounty. This makes The Hacker News headline from last Saturday all the more intriguing. Their headline was "Claude Opus 5 Helped Researchers Take Over OpenAI Staff Accounts via Chained Flaws." And there's a lot of meat in this. There's a lot for us to talk about here.

So here's what The Hacker News reported. They said: "Three researchers at the security firm Hacktron" - you've got to love that. So it sort of suggests maybe automated hacking, and that's exactly what it is. "The security firm Hacktron used Anthropic's Claude Opus 5 to chain two flaws and take over the ChatGPT and Codex accounts of several OpenAI employees, then reach an internal OpenAI code repository." In other words, a total breach of OpenAI's security leading to full access to OpenAI's proprietary internal code repository. Yikes.

The Hacker News continues, writing: "The chain began with a bug in the software that runs OpenAI's public help forum and moved through a weakness in OpenAI's own login system. But this was security research, not a real-world attack. The team reported the flaws to OpenAI, proved their access with a deliberately harmless pull request, then stopped. From the first look they made, obtaining that full internal access took under 72 hours.

"OpenAI confirmed a fix about 14 hours after the report, according to Hacktron, and on September 1st paid the team a $6,500 bounty. OpenAI said the award 'recognizes the OpenAI-side finding, not the actions against Discourse,' which is the open-source software that runs the forum. Testing the forum itself..."

Leo: That's the ones we use, by the way.

Steve: I know. Testing, yeah. "Testing the forum itself was out" - it turns out it was an image display bug. But we'll get there in a second. "Testing the forum itself was outside its bug bounty program." Right, because OpenAI is only talking about their side of it.

They said: "OpenAI has not publicly described the login flaw, and it confirmed the finding through that fix and payment rather than by detailing the account takeovers. Hacktron, which describes itself as an AI-assisted security research firm, was careful about what it did and did not do. When one employee's Codex link to OpenAI's code on GitHub was opened, it triggered a single pull request in the internal repository. It did not read any source code, merge, ship anything, or touch customer data.

"What the chain could have reached," they wrote, "was far larger. Because staff connects other services to ChatGPT and Codex, the team said the same access could in theory have extended to tools such as GitHub, Slack, and email. That wider reach was possible, but not used.

"The reason a bug in a public forum could reach staff accounts lies in OpenAI's login system, not in the forum software." In other words, not in Discourse. They said: "OpenAI's forum offers a 'Sign in with OpenAI' option, the same single sign-on that staff uses elsewhere." So thus a collision in single-sign-on overlap.

The article said: "Once the researchers took control of the forum server, the shared login let them take over the ChatGPT and Codex accounts of forum members who worked at OpenAI. The victims did not have to do anything." Again, it was just a credential overlap.

"Hacktron said this was an OpenAI identity problem, not a flaw in the forum software. Any first- or third-party service using the same sign-on could have granted the same access. The way in was an image bug. The forum runs on Discourse, and Discourse passes uploaded HEIC and HEIF (often verbalized as 'heef') images to a tool called ImageMagick" - very popular image rendering tool - "which uses the libheif library to read them. A flaw in libheif let a specially crafted image corrupt the forum server's memory."

Now, here's where it gets cool, Leo. "Discourse's advisory rates the result as remote code execution, scores it 8.8 out of 10, and tracks it as CVE-2026-32882." So Discourse should be updated, everybody, in order to get rid of this libheif exploit which is now known publicly. "The public record for the flaw itself is narrower. In libheif's own advisory and in national vulnerability databases, that CVE ending in 32882 is an out-of-bounds read that can crash the software or leak nearby memory, not a direct code-execution bug."

Here we go. "That leaked memory helps defeat a common protection called Address Space Layout Randomization. The researchers say they combined libheif's memory bugs, with the AI's help, to turn the crash into working code execution on the forum server. Upstream, the flaw was fixed in libheif 1.22.0 in May of 2026." So everyone will want to make sure that they're running an instance of ImageMagick that incorporates libheif with 1.22.0, which would have been sometime in May.

"That fix existed" - which is interesting - "months before the test. But the forum's server image, built on the Debian 12 Linux distribution, still shipped the old, unpatched libheif, version 1.19.7, when the researchers looked in July. The fix and its CVE were already public, but Debian had not yet included them in the packaged version the forum used.

"If you run your own Discourse server, this part affects you directly," they wrote. "Rebuild on the latest image to get the patched libheif because a web-interface update alone may not replace the old library. Sites hosted by Discourse were already patched, and the fixed self-hosted releases are 2026.7.0, 2026.6.1, 2026.5.2, and 2026.1.6.

"So, how did the researchers used AI? They used AI to do the hard part. They first" - get this, Leo. "They first tried Claude Opus 4.8, which struggled over several sessions to build a working exploit once ASLR was enabled." I'll have a lot to say about this in a minute. "Anthropic released its next model, Claude Opus 5, on the evening of July 24th, and in a fresh session it produced a working code exploit within hours."

> **Leo:** Wow.

Steve: So here we have Opus 4.8 can't do it, no matter how many times they try. Anthropic updates Claude to Opus 5, bang, cuts through it like butter. They said: "Opus 5 shipped with safeguards" - okay, here comes more. "Opus 5 shipped with safeguards meant to stop it from writing exploit code for real targets." Which it did. "The researchers got around them by pointing the model at their own test server, disguised as a capture-the-flag practice target, then letting it run in an automated loop. Even so, they say, the work was not hands-off. Skilled human direction still mattered, and this was not automated hacking with no one at the controls.

"The case fits what researchers and AI companies have described this year: capable AI models are sharply cutting the time and skill that serious offensive work used to require. Anthropic has reported that criminal and state-backed groups are already using its Claude models to run real intrusions, not just to ask and answer questions.

"OpenAI was one target in a wider project Hacktron calls HEIF Heist. Over about two months, the team says, it found the same class of image-decoding flaws in software used by other large companies" - right, because everyone's using this - "at a total cost of under $3,000 in their AI usage. It links the campaign to reported bugs in Slack, Meta's products, GitHub Enterprise, and web frameworks such as Next.js.

"Those broader claims are backed unevenly. The Next.js flaw is confirmed in Vercel's own advisory, and libheif's maintainers confirmed a working code-execution exploit for the bug tied to Meta. The

wider claim of code execution across many applications has not been independently confirmed, a limit The Hacker News noted when it first covered the Next.js flaw in August.

"The wider campaign used a different model, OpenAI's own GPT-5.6 Sol, for cases where the team knew nothing about the target in advance. Only one company, Shopify, appears to have noticed the activity, the researchers say, even though its image processors crashed repeatedly under thousands of test uploads." In other words, many companies don't notice that somebody is hacking their systems. Only Shopify said, hey, what the hell's going on here? Most companies were completely oblivious to it. Again, that's a problem.

"The Hacker News has contacted Hacktron with questions about how the forum code execution was achieved and about the scope of the account access.

"So what should users do? The bigger lessons go beyond Discourse. If your service accepts user images and reads HEIC, HEIF, or AVIF files through libheif, an old build could be exposed. And if a public, lower-trust service shares your single sign-on with internal tools" - which was the mistake OpenAI had - "a break-in on that service can become a break-in everywhere the same login reaches.

"Update libheif to the latest security release (1.23.4 as of early September 2026) or to your distribution's patched build. Where you do not need it, turn off decoding of untrusted HEIF and AVIF images, or run image processing inside a locked-down sandbox. Limit which services your single sign-on trusts, and require a fresh identity check before sensitive actions rather than trusting an existing session." Again, it's that common session reuse problem is where this bit OpenAI.

They finish: "There is no sign the OpenAI flaw was used against anyone in the real world. As of mid-September 2026, it was not on the U.S. government's list of vulnerabilities known to be exploited" - CISA's KEV - "although that list is not proof either way. What the available reports do not settle is whether an organization that has already patched should still check for earlier access; on that point, all sources are silent."

Okay. So we learned a number of very interesting things from this reporting. It was, I think, really interesting, as I said, that whereas Claude Opus 4.8 worked and worked, but was unable to develop an exploit, simply upgrading to Claude Opus 5 handed the would-be attackers a working exploit within hours. This pretty clearly demonstrates the rapidly evolving effective strength of available AI.

The next chilling aspect revealed by this report is the successful ease with which the attackers were able to trick both Claude Opus 4.8 and 5 into bypassing their own guardrails to develop a working exploit for them. They used, essentially, the "Oh, Claude, it's okay. We're a cybersecurity firm, and we're just wishing to determine whether the trouble we've identified in the widely used libheif library could be weaponized in the presence of ASLR. So we just need you to try to do that for us on our own internal test server. Would you do that for us, pretty please?" As I have said from the start, today's AI technology has always felt very slippery and extremely difficult to control.

And finally, the most worrisome evidence we obtain from this is that the work that Claude Opus 5 performed for these researchers was truly significant. Just to remind everyone, Address Space Layout Randomization (ASLR) was added to systems as a means of thwarting attackers who managed to obtain execution inside a protected code space such as an operating system kernel. Before Address Space Layout Randomization was added, a system would always load its many various modules, essentially stacking them, in the same order, inside the operating system's RAM memory. This meant that an attacker could count on what was located where when they wished to use, for example, return oriented programming (ROP) to knit together the privilege changes they needed to make by using the operating system's own existing code at the ends of existing subroutines. They would jump near the end of a known subroutine.

They would do a little something that they needed before it returned. And when it returned, it returned to them, being a subroutine. So ASLR was added to make doing this far, far more difficult. As the name says, Address Space Layout Randomization deliberately does NOT load all of the system's modules the same way each time. In fact, it goes to extreme measures to scramble up the loading so that no code an attacker might manage to get running inside an operating system kernel will have any idea where anything else is located. ASLR, consequently, has proven to be an extremely effective and essentially critical attack success mitigation technique.

The technique is so difficult to bypass that Claude Opus 4.8 was unable to succeed. It stopped it cold. But then along comes Claude Opus 5, which reportedly cut right through ASLR. Understand how bad that is. ASLR bypasses have been known in the past because it is, at best, a mitigation, a strong mitigation, but still just a mitigation. It is not, never has been, and cannot be a total prevention. It makes the attacker's job far more difficult, but it is unable to make it impossible. Until now, ASLR has meant that only the upper echelon of elite hackers, you know, these Hacktron people couldn't. Only the upper echelon of elite hackers could find their way past it. Claude Opus 5 just changed that. Now, all any script kiddie needs to do is ask for passage no elite hacking skill required.

So I am very glad that there are now white hat hacking firms like Hacktron which have decided to leverage the power of AI to find and report remotely exploitable vulnerabilities and collect bounties. That's great. You know, we need more groups like them. It would be terrific if they could be granted fully cyber-capable access to Anthropic's best models without guardrails so that they don't need to try to trick Claude to do what they would want to do in order to protect firms from themselves. These are the sorts of people we want to up-arm in what is going to be something of a little bit of a cyberwar here for a while.

Leo: I like that verb "up-arm." That's new to me. I like it.

Steve: Good.

Leo: Up-arm it.

Steve: Up-arm it.

Leo: Get prepared. Arm your ups.

Steve: Okay. So also last Thursday, The Hacker News carried the report of a maximum severity, and you never want to read this about something that Cisco is offering, a maximum severity CVSS 10.0, which we know is incredibly difficult to get. I mean, basically, if you have a vulnerability of 10.0, the device is saying, come on, come on in. Just, you know, come in, look around. Take over my enterprise that purchased me.

It is remotely exploitable to give anyone who wishes root access and full control over an affected Cisco device. Starting off the article, The Hacker News wrote: "Cisco has warned of a fresh maximum-severity security flaw impacting Identity Services Engine (ISE) that has come under active exploitation." So the word is out. "The vulnerability, tracked as CVE-2026-76460, with a CVSS of 10.0, could allow" - and we know now does, is, does, and will - "allow any unauthenticated remote attacker to bypass authentication. Cisco wrote: 'This vulnerability is due to insufficient authentication control' - which, yeah, no kidding - 'on an API endpoint. An attacker could exploit this vulnerability by sending a crafted request to an affected API endpoint. A successful exploit could allow the attacker to gain unauthorized access to the affected device by bypassing the web-based management interface.'"

Well, our listeners well know how much disdain I have for any and all publicly exposed web interfaces. It's almost to the point now where you deserve what you get if you expose a web interface. All of the evidence we've seen for years proves beyond any doubt that, for whatever reason, we just do not seem to know how to secure a web interface. We just don't. We can't, apparently. If these problems directly damaged Cisco, they might have removed those inherently insecure interfaces long ago. But since Cisco's users want them, and since it's they who are damaged by the endless ransomware extortion campaigns they enable, nothing ever changes. Even when Cisco comes under hopeful new management, as we covered last year sometime, that promises to put their customer's security ahead, even ahead of the customer's own convenience, well, that doesn't seem to change.

What has changed for the better is that, as I said at the top of the show, Cisco, like Microsoft, has apparently also discovered that AI can help them to finally root out these latent legacy problems. Last Wednesday, the day before that reporting, Cisco dropped what is for them a stunning ensemble of 77 patches for systems across their product line. I'm not going to go through it all, but I'll give you a sense for it. All the CVEs are of course 2026. So 2176 has a CVSS of 9.9; 20211, 9.1; 20307, 9.1. All multiple vulnerabilities in that ISE that could allow an authenticated, remote attacker to execute arbitrary commands on the underlying operating system of an affected device. To exploit these vulnerabilities, the attacker must have valid admin credentials. Unfortunately, if you have that, you can just take over the device, do whatever you want to outside of normal, you know, the normal controls.

Then we have 76423, carrying a CVSS of 10.0; 76424, 76425, 76426, 76427, 76428. Multiple vulnerabilities in ISE and Cisco ISE Passive Identity Connector (ISE-PIC) that could allow, and we now know does, a remote attacker to bypass authentication to the REST API, achieve remote code execution, perform SQL injection, and conduct XML External Entity injection attacks on an affected device.

CVE-20282 with a CVSS of 9.1; 20283; 20284. Multiple vulnerabilities in ISE [blah blah blah]. CVSS 9.1 and 9.1. Multiple vulnerabilities in ISE and ISE-PIC and so forth, you know, remote attackers perform command injection.

CVE-20322 carrying a CVSS of 9.9; 20325 with CVSS of 9.9; CVE-20326, CVSS of 9.8; 20360, 20361, 76409. Multiple vulnerabilities in Cisco Nexus Dashboard that could lead to command injection, authentication or authorization bypass, and information disclosure.

CVE-20130 with a CVSS of 10.0; 20192, CVSS of 10.0; CVE-20194, a CVSS of 9.1; then another 9.9 and 9.9). Multiple vulnerabilities in ISE [blah blah blah]. And it goes on like that for a total of 77. And they are all in the high nines of CVSS scores. We don't have to guess where these all came from; right? We see the pattern that the employees of massive legacy enterprises, Microsoft and Cisco come to mind, the employees are unable to clean up their own legacy code. But as I said at the top, when those cash-rich enterprises can pay to have AI do that for them, their code gets fixed. And that is nothing but really great news. So at this rate, as with Microsoft, Cisco's platforms may at last finally become securable, which would be wonderful.

Leo: Wouldn't that. I mean, they're everywhere; right? They're installed in everything.

Steve: Oh my god, yes. Cisco was first, and they've always been like the IBM; right?

Leo: Right.

Steve: You can't go wrong by ordering Cisco. Which is, you know, it's been the bane of many of their competitors who have said, you know, their stuff is not that great anymore. Look at what we have. But people like that dark green color, apparently, of the...

Leo: Well, and it's, you know, in their defense and Microsoft's defense, when you've been around forever and ever, your code base gets a little old and grungy and, you know. But this is good. We're getting this stuff fixed up now.

Steve: Yes. I salute them. If you cannot hire people to fix your legacy code, and apparently Microsoft and Cisco can't, then hire AI to do it. And it is doing it. I mean, this is a grand slam, 77 serious external exposure vulnerabilities in Cisco. What it does mean is you've got to get your gear updated because the bad guys are going to backtrack these vulnerabilities and attack you with them. So really, really, really update Cisco. If any of our listeners have any responsibility for Cisco gear on the edge, make sure that it's updated. And Leo?

Leo: Yes.

Steve: It's time for our last break, and then we're going to take a look at, speaking of a super rich company that is spending a lot on AI to fix their problems, it turns out it's not all good news. The Mega Patch Tuesday Fallout...

Leo: Uh-oh.

Steve: ...is our final topic.

Leo: There was a little bit, yes. But we'll talk. It's good. Okay.

Steve: More than you know.

Leo: Oh. Hmm. Hmm.

Steve: Yeah, really bad.

Leo: Well, you know, that's why people come here. They want to hear the bad news with the good news. In fact, it's mostly bad news. But if you are responsible for keeping your company's systems running or protecting your security, this is the news you need to know; right?

Steve: Last week we noted that, while it was super-spiffy-terrific that Microsoft has now leveraged AI to fix things left and right with each month breaking all previous historic records, there was also some consequential downside in the form of the massive burden that all of these patches would have on enterprise IT staff who now needed to make sure that applying these fixes did not break anything that their enterprise might depend upon.

And despite the need to proceed with caution, just days before the release of these patches, which was now two weeks ago, or three weeks ago, Microsoft publicly warned their customers to apply these fixes without delay with the argument that AI-assisted exploit development was now shortening the time window to exploit. So it was somewhat ironic that it was Microsoft, it turns out, that needed to quickly pivot into "emergency mode" to address the many widespread catastrophes that were created by Tuesday's updates.

What happened? Well, there's the RDP Hang: Within 48 hours, admins began reporting that RDS (Remote Desktop Services) on patched Windows servers would run for a few hours and then die. New sessions would hang at the "Connecting..." dialog or at the "Please wait for the Remote Desktop Configuration" screen. And at the time of the collapse, any existing sessions that were in place were unable to disconnect or log off. It was a mess because so many organizations have become quite dependent upon Remote Desktop. Which of course Microsoft has been pushing as part of their whole virtualization move.

But the trouble, it turns out, was not limited to RDP. Also when this occurred, MMC (the Microsoft Management Console), the RDS Licensing Diagnoser, even Windows File Explorer and Windows Update settings page would all become unresponsive on the affected machine. Admins were forced to hard-reset production servers during the day. And the blast radius of this one was pretty significant: Server 2012, 2012 R2, 2016, 2019, 2022, 2025, plus Windows 10 21H2/22H2 and Windows 11 24H2/25H2/26H1. Like, frankly, the server I'm still running is 2008. It was spared because it was older than...

Leo: I forgot it existed. What are you talking about?

Steve: That's right. Although I won't be - I'll be up to Server 2025 as soon as I bring the new hardware online.

Leo: Oh, good.

Steve: But I'm not there yet. So anyway, Microsoft's reaction was to offer what is known as a "known issue rollback," which, you know, they've had to do in the past. Which is, again, this is why enterprise IT is a little, you know, careful with these Windows Patch Tuesdays. The "known issue rollback" could be administered via Group Policy. And then, finally, on the 14th of September, six days after Patch Tuesday, so it was - actually it was Monday, the day before last week's podcast - Microsoft pushed seven individual out-of-band updates out to cover all troubled target machines. And for the various servers, unfortunately, on the server side, those updates were catalog-only, so they do not come through Windows Update, and admins were required to get them manually. So that was a problem. But arguably not the worst.

There was the Domain Authentication Mess: In some instances - and again, the source of this one was really sort of interesting. Windows 11 users were prevented from logging in with valid domain credentials after installing the September 2026 security updates, you know, Patch Tuesday for this month. Microsoft has been gradually promoting a feature called Machine Identity Isolation. It tightens how a machine account goes about proving itself over the Netlogon secure channel. Until two Tuesdays ago, that is, before Patch Tuesday, that feature, machine identity isolation, had been present, but it was set to a "configured but not enforced" state. You know, this is sort of the way Microsoft creeps these things out, you know, Microsoft and others, where they sort of incrementally roll them out, but they don't commit to them.

It's the way XP originally had a Windows Firewall. It was the first version that had a firewall. But it was there, but it wasn't enabled by default. It wasn't until Service Pack 3, I think, was it 2 or 3 that turned on the firewall by default.

Anyway, so same thing here. Machine Identity Isolation configured, but not enforced. Well, so you can guess what comes next. September's updates flipped it into its configured state with enforcement mode on machines where it had been previously only configured. The only problem was that it turns out it's only supported when the domain is at Windows Server 2025 Domain Functional Level, which it turns out is a small minority of the world.

Leo: Oh, geez.

Steve: That's right. As a result, anyone, any enterprises running Server 2019 or 2022 Domain Controllers which, again, is way the majority, most of the world they found that their clients could no longer authenticate to their domains. Whoopsie.

Leo: That kind of puts pasting in Excel in the shadows.

Steve: Yes, it does, yes.

Leo: Yikes.

Steve: Users receive the disturbing and clear message: "The trust relationship between this workstation and the primary domain failed." Now, since cached credentials still work offline, users could often log in at home onto their machine, but then find themselves locked out of the corporate

network. Windows' Credential Guard machine-account authentication also broke, and VPN-connected devices were especially prone to trouble because they also use the same trust relationship. So it was and still is a mess. Microsoft has acknowledged the trouble last Thursday and offered a workaround, but not a patch.

The workaround - get this - is to disable the Machine Identity Isolation feature by whatever means enabled it, which might be Intune, Group Policy, or the registry. So you set the MachineIdentityIsolation value to 0, then reboot, and then you have to execute a powershell command. Unfortunately, Microsoft's documentation warns that enabling Machine Identity Isolation in enforcement mode, which is what they did on Patch Tuesday, then subsequently disabling it, which is what they are now saying is necessary, will break domain authentication and require the device to be unjoined and rejoined to the Windows domain. I'm sure this was a security feature which they built in, you know, to be extra safe; but that does make recovery more burdensome.

One solution is to run a powershell command to repair the broken secure channel. If anyone out there is being affected by this, I've got the powershell command in the show notes. It's Test-ComputerSecureChannel -Repair -Credential (Get-Credential). And that will unjoin and rejoin that machine to the domain. And then you can then log in. So you can imagine how many enterprises got bit hard by this and took away the lesson, ooh, we're really - it doesn't matter how much Microsoft tells us we need to update. This is going to be a problem.

And then of course I actually heard from several of our own listeners who are responsible for users within their organization about Excel's copy and paste being broken. The update that was responsible was KB5002914. The good news is that that update resolved 29 different Excel vulnerabilities which included a number of which could be used for serious remote code execution exploits.

Leo: Well, that's good, yeah.

Steve: Yes. So there's that. But in the process it also broke our good old copy and paste. The thing that caused so much trouble was that the copy and paste failed silently. Microsoft wrote: "Although users try to paste content, the source remains selected and the destination is unmodified." So no error, no dialog. CTRL+C followed by CTRL+V, you know, copy and paste, does nothing. Autofill doesn't work. Dragging the fill handle doesn't work. Fill Series doesn't work. Dragging cells to move them doesn't work. Interestingly, Office 365 users were not affected, only those using Office 2016, 2019, Office 2024, or the LTSC, you know, the long-term servicing channel 2021, they were affected. And of course since many people spend significant time working in Excel, this was all debilitating for them, you know, not what you want from a Windows Patch Tuesday.

Then there was USB Audio: USB Audio Class 1.0 devices all stopped working with Device Manager complaining: "This device cannot start and giving a Code 10." The result was silence, with volume controls unresponsive, the sound settings page sometimes crashing. In a related symptom, multichannel devices lost their 8-channel and 3D audio modes, although stereo kept working.

Then there was the WinRE partition too small bugs: The Tuesday update included a Secure Boot certificate update which was larger than the previous certificate. That required the need to resize the Windows recovery partition. Unfortunately, systems whose WinRE partition was smaller than Microsoft's recommended 750MB, which happens to include an enormous number of OEM images that do not wish to waste so much space because they would rather give it to their users, that would all fail the resize and often then fail to boot. Dell's OptiPlex machines were widely reported victims of this. It was a mess.

Then there's the Code Integrity Failure: When attempting to boot, HP laptops in particular were hitting a stop code of 0xc0430001, "Code Integrity failed to initialize," and then receiving a black screen or an infinite restart loop because of a Secure Boot file mismatch. The popular workaround that's been suggested is to disable Secure Boot in the BIOS, but of course that means retrieving the system's BitLocker recovery key first.

There have also been reports of KERNEL_SECURITY_CHECK_FAILURE boot loops on older hardware. In these cases it was not possible to simply uninstall the update because the machine would not boot far enough to allow the user to do it. Again, another huge mess.

And then there was the Windows Desktop Explorer.exe, like the actual Explorer.exe, which is the Windows Desktop. It was crashing on any VDI systems, you know, Microsoft's Virtual Desktop Infrastructure. So all the enterprises using Windows VDI through Citrix UPM, FSLogix, VMware Horizon, Liquidware ProfileUnity, they all virtualize the desktop with remote servers reporting that Windows Explorer desktop failed to start or was crashing immediately after sign-in, leaving a blank screen with no taskbar. A Microsoft engineer confirmed they're investigating and asked for memory dumps please, but there's no root cause and no fix. The workaround is restarting Explorer from the Windows Task Manager, or just use a fresh profile. But of course then you've got a whole blank Windows. For an enterprise with thousands of pooled desktops, this is another nightmare.

And, believe it or not, there's more. I'm not going to go into such detail, but AMD Radeon video, many of their cards suffered black screens, driver timeouts, and hard freezes. There was the "BitLocker Event ID 24641" which recurred after every reboot. The message that users received was "An unexpected error was encountered attempting to retrieve the BitLocker volume master key during restart." Now, that's not what you want to see. However, the drives remained accessible, so it appeared to be some sort of worrisome cosmetic glitch. But it upset lots of people who were wondering what was going on.

Hyper-V and Plan9 shared folders stopped working. Always On VPN with IKEv2 certificate-based VPN, those connections began failing and dropping immediately. There was a later fix to that using KB5129195. And finally, update installation failures. Windows 11 LTSC, the long-term servicing channel 2024, and various Dell notebooks would not accept Windows updates. They would roll them back at the first reboot. And then it was necessary to reinstall them by hand.

So I wanted to take the time to walk through these rather than just waving my hand and saying there were lots of problems two weeks ago because it's important to highlight the tight spot this puts enterprise IT staff in, and the spot they may be in next month, two weeks from now. Microsoft had said "This is a really important biggie, and due to the new speed of AI-driven attacks, everyone should patch within three days." And this admonishment was offered, as I said, shortly before the delivery of this massive patch update that did in fact wreak havoc within the networks of many enterprise users.

So what do we do about this? I don't think there's anything we can do. Overall, I think Microsoft did the right thing. Their entire product suite is measurably more secure today than it was two weeks ago. Nearly a thousand patches worth. But they need to learn from this to determine why they did not detect, you know, these update side effects. You know, maybe they were in a big hurry. Maybe junior people are asking AI to fix problems, and they're not vetting them enough. I mean, why were all these things, how did they go undetected?

But even if the same thing happens next month, it's unfortunate, and I feel for the enterprise IT guys who are going to be terrified to apply next Tuesday's patches. We don't know how many there are going to be. Only Microsoft knows, and we don't know what quality they're going to be. But I'm afraid we're going to have to go through some pain like this to finally get to a Windows which not only works for everybody, which of course is always the goal, but is also secure.

Leo: There were plenty, god knows. Geez Louise.

Steve: There were plenty. And again, I don't know what enterprise does. I mean, they cannot afford to apply these except to do it in their own test environment.

Leo: Yeah.

Steve: They're going to have to test them and see if, like, what breaks. But I'm afraid breakage is the way we get there because, you know, who knows what next month is going to look like? That's two weeks away. And this is going to happen again.

Leo: Oh, man. I do not envy the job of our IT professionals who listen to this show. You guys...

Steve: And you don't want to not patch.

Leo: No.

Steve: Because then you could be victim to...

Leo: Now known flaws.

Steve: ...flaws which are now known, exactly.

Leo: They're out there. Yeah, better to fix them, I guess. The problem is there's such an onslaught, this is what I was saying two weeks ago, that Microsoft probably doesn't have time or the ability to test them. Not as thoroughly as they'd like.

Steve: There's probably an urgency on their part to get them shipped.

Leo: I mean...

Steve: And unfortunately when that happens, this happens.

Leo: A few years ago, I will have to ask Paul about this, Paul and Richard tomorrow on Windows Weekly, but a few years ago they got rid of their testing team, which was maybe a bit of a...

Steve: We talked about it here also, yeah. Yeah, we'll just let our customers be our...

Leo: I mean, right now you need, you want a thousand people. You want somebody on every possible device. You know, you want every computer ever made, every server...

Steve: Unfortunately, you cannot scale that up overnight. And that's what they've had to do.

Leo: And I think, you know, one of the reasons I kind of laugh when I hear of these doomers talking about how AI's going to kill us all - and you'll see this when you start playing with local AI. So AI is so smart, it found the flaws, but it didn't find all the side effects. If it was so smart, you'd think it would have noticed that this is going to break paste. And this has been my experience. They do amazing things, but there are gaps. There are places where they're incredibly stupid.

Steve: Well, Leo, because they don't actually understand anything. There is no understanding. It is astonishing that just language statistics is able to give us this. That's all it is. It is language statistics. We have so much knowledge stored in language that you can query it. You can squeeze it out. But of course it doesn't understand anything.

Leo: So I'm having the same experience with my own coding. I'm really very, very careful. I have auditors. I have, you know, some really elaborate checks and balances. And yet, man, I thought, oh, it's really good. But the other day I was kind of dejected. I thought, I've got this ad sales system down. Everything's working beautifully. All the numbers work out. It looks beautiful. I said, "Lisa, we're ready." And I showed it to her. And the very first thing she does, it doesn't crash, it says, oh, you can't save that. There's another user. But there's no other user. What are you, nuts? It's just us. And it's like, ah. So close. And so it's frustrating because, yeah, they're not genius engineers. Even genius engineers make that kind of mistake. But these guys...

Steve: Well, what we don't yet know is whether AI code slop, or to what degree AI code slop is going to be a problem. Are there, you know, I just enumerated the major problems that were created by Microsoft dropping nearly a thousand patches.

Leo: Right.

Steve: We don't know if there are subtler problems like maybe when you - now we know...

Leo: I guarantee you there are subtler problems. Guarantee you. You know, paste the letter Q into Excel, and the whole thing goes kablooie. I guarantee you.

Steve: Right.

Leo: Because you can't test everything.

Steve: Right.

Leo: It seems to me in theory - I don't know, you're a very accomplished coder. I guess my - Windows is just too big, and Office is too big. But in theory it's deterministic. You know where everything is. You don't, though. Wire it up and what every - the impact will be of every change you make. It should all be cause and effect.

Steve: Well, I'm still of the opinion that we could have a much better coding system because this is just using sample code from the Internet to knit together solutions. It's, I mean, code is code. It has laws and rules. And, I mean...

Leo: It's all deterministic, ultimately.

Steve: I mentioned to you, I think it was before we were recording, that all of this has been done with a linguistic model. There are other models now underway. There are world models. There are models for specific areas, like Microsoft apparently is training up a materials model for some reason. So it won't be language. It'll be something else. And it would be entirely possible to train, I think, somehow, a code model that is not about language, but is actually about code. So...

Leo: You know, there's a new model out there which you will be able to play with once you get your Spark plugged in called JEV, J-E-V based, it's Jevons Paradox, you know. And it's a classifier. It's not an LLM. It doesn't do prose. Doesn't do language. And it's very fast. And it's very interesting. And I'm actually starting - I have about 11 tests going on in different things

because it is a very fast and effective way to say this or that. And you give it the criteria. It gives your percentage match. And it's really good. And it isn't language.

Steve: This is that branch that wonders if large language models are not compressors. Because Gzip is basically a statistical compression. And so it turns out that you can do some odd things. It's sort of a branch, what you're talking about. So I have sort of run across that. But I haven't played with it yet.

Leo: It's some very interesting stuff. And because it's so fast, it's very cheap. They don't even charge you for tokens out, just tokens in. It's really - this is the good news. We're at the very beginning of a revolution.

Steve: Leo, that's what I keep saying is that, yes. This is fun. We are riding a tidal wave. And nothing we know today is true 60 days from now.

Leo: Which is why I'm really, really glad that you got a Spark because I want you to start digging deep into this. I'm very interested.

Steve: Well, I want to accelerate my app development because it's dumb that it takes me years to create an app.

Leo: Yeah.

Steve: So I'm - the only thing I would ever have it do would be to build my UI. But I spend an awful lot of time moving buttons around and tweaking margins and centering things and doing stuff that's dumb for me to spend time on.

Leo: Yeah, that an AI could do very well.

Steve: So having a core, my own MASM core in the back end and a clean interface API between it and the UI, then I'm never - I will always be using cloud, agentic cloud coding to code the UI because I'm only going to go for the best, and local is never going to be as good as the cloud.

Leo: No, it's true, yeah.

Steve: So at least for the foreseeable future. I do foresee a day when Microsoft is offering enterprises an AI server. There will be, you know, you will be able to - there will be a Microsoft AI server.

Leo: Oh, for sure.

Steve: And the other thing that is weird, I shot you a note about it the other day, is to remind ourselves that none of this AI is training. I mean, all we're doing is building big prompts around a static AI. It's wrong that an AI that a corporation uses wouldn't learn about like the right things about what the corporation is doing. It's just, you know, we're just building really fancy scaffolding around static AI. So we're just...

Leo: Just at the beginning, yeah.

Steve: You know, baby steps.

Leo: Well, I'm excited. And I'm so glad you're here to guide us through it. Steve Gibson's at GRC.com, the Gibson Research Corporation. That's where you'll find SpinRite, the world's best mass storage maintenance recovery and performance enhancing utility. It's a must-have, 6.1 the current version. You can go there and get it. If you already have it, you can get the upgrade for free.

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/