How Worried Should We Be?

Description: Muse has a bad zero-day. The regularity of "Irregular." More rogue OpenAI breaches. The Seven Deadly Sins (TSDS) hacker group. Liquefied Natural Gas (LNG) cargo ship hacked. The FBI offended ShinyHunters' delicate sensibilities. Canonical switches to an every-two-weeks release cadence. Axios AI 101: AI Explainer for normal people. How worried should we be?

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1098.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1098-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. We're going to talk about, well, the new AI agents like Muse, that are a little bit spooky. A new hacker group called The Seven Deadly Sins, you can only imagine. And one of the worst breaches in history. The FBI should apologize, says Steve. Stay tuned. Security Now! is next.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1098, recorded Tuesday, September 29th, 2026: How Worried Should We Be?

It's time for Security Now!, the show where we cover the latest in security, privacy, and all that jazz with the one and only, the legend, Mr. Steve Gibson, the guy who has been doing this longer than anybody, coined the phrase "malware" or "spyware," wrote the first spyware tool. We don't worry about spyware as much anymore, do we. It's just endemic. You know you're being spied on everywhere you go, Steve.

Steve Gibson: Right. We gave up. We lost that race.

Leo: We gave up. We gave up.

Steve: It was like, you know, yeah. And also we'll hear me later this podcast remind everyone that, if it doesn't make money for the bad guys, it doesn't happen.

Leo: Right.

Steve: That's one of the reasons that I'm not that worried. Today's topic, today's title of our podcast is "How Worried Should We Be?"

Leo: Oh. How worried should we be?

Steve: And I was talking to you before we began recording, and I noted that just the title indicates what we'll be talking about. That is to say, like, about what? Well, no one wonders today what it is.

Leo: We know what we're worried about, yeah.

Steve: Yes. It's the only thing going on right now. So, and you also noted that you get some feedback from people saying would you stop talking about AI? I want to say that the people who email me hear that I'm self-conscious about us spending so much time, and they say don't worry about it. We want to know what you think about AI because it often differs from everything we hear elsewhere.

Leo: That's a good point. You have a unique perspective on it, yeah.

Steve: Yeah, I mean, who was the first person you ever heard say it's going to be very good with code, and you also heard say it's going to be very difficult to control this?

Leo: Yeah.

Steve: Both those things I said, like, years ago. And it's playing out now. So anyway, how worried should we be for Episode 1098? Will we get to 1200? Well, I don't know. No. If you want the short version, it's not very.

Leo: Good.

Steve: But we're going to take a long way to get to that conclusion because there are some interesting things I think that we have to talk about and look at along the way. We're going to talk about Muse's very bad zero-day. And actually there's another one, a VM breakout that also just happened. So Muse is having some troubles on the security front, of course Meta's new agent. And just hours ago OpenAI introduced Dots, which is going to be their agentic take. We've got somebody else, as I sent a note to you a couple days ago, Leo, that it was nice to see somebody else noticing, as you and I had, that Irregular, the company Irregular, was a common factor in many of these breakout stories. We've got more rogue OpenAI breaches.

We're going to introduce The Seven Deadly Sins (TSDS), The Seven Deadly Sins, TSDS hacker group. An interesting hack of an LNG, a Liquefied Natural Gas cargo ship. Actually, this was the third in a series, and it's got officials worried. We've got the well-known ShinyHunters group we've been talking about for quite a while, which we know is an amalgam of several other groups which have gotten together under that moniker, hacked into the FBI's site. And I have a somewhat controversial suggestion about how the FBI should handle that.

We've also got some news about the people behind the Internet's number one Linux distro, which of course is Ubuntu. Canonical switching to an every-two-weeks release cadence. And of course we know why. I found an amazing AI explainer which Jim VandeHei with Axios produced. I gave it a GRC shortcut because it is so good. It is what all of our listeners can give to their friends who want to understand what in the world is going on. It's incredibly approachable. And then we're going to talk about this question. I think we can answer it now. At least, you know, evidence based, how worried should we be, actually?

Leo: Good, good.

Steve: Because the world's gone insane over all this.

Leo: Good. 10%? That always cracks me up. Like you could calculate...

Steve: I know. That is so bogus, so oh my...

Leo: They should say 11.8%, and then it would be like, oh, it must be math. Science.

Steve: And unfortunately, you couldn't get a better soundbite to get repeated over and over and over.

Leo: Oh, yeah. Oh, yeah. Everybody understands it, 10%.

Steve: Anyway, we're going to have I think an interesting podcast for everybody.

Leo: Can't wait.

Steve: For a change. Okay. So we do have a Picture of the Week.

Leo: Oh.

Steve: And it's fun. I got the picture, and I thought, okay, how could I caption this? And I thought, okay, those fluorescent light bulbs in the Public Storage alley keep being stolen, Leo.

Leo: The fluorescent light bulbs.

Steve: The fluorescent light bulbs in the Public Storage alley keep getting stolen.

Leo: All right. I haven't looked at this. I'm going...

Steve: What are you going to do about that? How do you solve these...

Leo: I don't know what this is referring to. Let me scroll up.

Steve: How do you solve that problem?

Leo: No. No. You'd better describe this one.

Steve: So the fluorescent light bulbs often take the shape of a long tube which is - sort of spirals out and then crosses over at the top and then spirals...

Leo: Yu don't see these much anymore. They've got, you know, there were a few years where this was everywhere.

Steve: Right.

Leo: Now it's all LEDs. But those are expensive, those fluorescent...

Steve: They're expensive. And, you know, in a public facility where - normally I think they have cameras, security cameras monitoring everything these days. But how would you keep that from being stolen?

Leo: Easily done.

Steve: Well, somebody came up with a clever idea, and it's another one in our series of fun ways to use a padlock. They looped a chain through the loop at the end of the fluorescent light, through a cross-member above the light, and padlocked it. So you could, you know, a bad guy could destroy the light, but then they're not getting any benefit from stealing it.

Leo: True, true.

Steve: Basically this completely solves the problem. I mean, it's going to end theft of otherwise easily, you know, exposed screw-in base fluorescent lights, which people might think, hey, you know, mine burned out at home. I'm going to just steal this one from a Public Storage place.

Leo: Sure, why wouldn't you?

Steve: That's right.

Leo: Absolutely.

Steve: Okay. So last week we briefly acknowledged Muse, which is Meta's, like, consumer-oriented Agentic AI Assistant.

Leo: And I have been using it like crazy, by the way.

Steve: And have you been?

Leo: Well, I'm going to just tell you what it did. I accidentally, my agent, my AI accidentally took my entire computer offline, which meant I had no access to my agents. And one of the problems, we've been talking about this, letting an AI be your sysadmin, is you gradually forget how everything works; right? All of a sudden I sit down at my computer, and it's not online. I can't access Claude. I can't access Codex. I can't access Hermes. And I don't know what's wrong.

Steve: Right.

Leo: I have no idea. I don't know why I thought of this, but I went to Muse, and I said, "Hey, you've got to help me, man. I can't get online." Now, fortunately I had given Muse access to the tail net so that I could use it outside of the house, and it would reach into my home network. It said, oh, I know what's happened. Hermes pointed my computer at an exit node on one of these little broom devices, it's a little travel router, and then broke the travel router, taking us offline. I said, "You committed suicide. What did you do? What did you do?" Fortunately, Muse, because it

could get to the tail net, said oh, I see what's happening, your framework's pointing to an exit node that no longer exists. It repointed it. Everything came back. So thank you, Muse.

Steve: Wow.

Leo: But this argues for giving Muse all the permissions. Which you're going to tell me is a very dumb thing to do.

Steve: This is the - what I predict with all of this agentic consumer-grade AI is...

Leo: Oh, everybody's doing it.

Steve: They are. And there's of course already one famous anecdote of some guy who asked Muse to sell something for him, and Muse sold it for a lower price to another person than this guy wanted to sell it for, and gave the other person the owner's home address. And this person comes and knocks at the front door. And the guy who was using Muse said, "What? What do you mean you sold it? How much for?" Anyway, we're going to see a bunch of stuff go wrong. It's just inevitable. I think it is probably going to be the place that we spin, that the industry spins on, longer than anything else is the fundamental uncontrollability or unpredictability of agentic AI. And Leo, just anecdotally, we've heard from you, like your agents stop. They just stop working.

Leo: Oh, yeah. Oh, yeah.

Steve: They go, well, we're waiting for you to click your heels three times. What? What are you talking about?

Leo: Yes, exactly. Oh, you wanted us to keep working. Oh, yeah, that's right, you told us that; didn't you.

Steve: Oh, okay.

Leo: Never mind, yeah.

Steve: So oh, my god. Anyway, so...

Leo: They have a mind of their own, that's very...

Steve: Well, and that's the problem. We want them to have a mind of their own. We have given them a mind of their own.

Leo: True.

Steve: Because that's how they're useful to us.

Leo: Yes.

Steve: And this is the great dilemma is that, you know, with great power comes great responsibility. Unfortunately, they're irresponsible. And we're trying to teach them responsibility, and I don't know how well that's going to work, or how quickly we're going to be able to. But boy, I said to Lorrie this morning, I was just shaking my head because I was reading the news release of OpenAI's. Just this morning OpenAI released Dots, D-O-T-S. Dots is their consumer...

Leo: Same thing, it's just like Muse.

Steve: ...focused agentic AI, yes. And I said to Lorrie, I said, "What a wonderful time to be alive." I mean, this is just, you know, it's chaos. And but cool stuff, you know. It's tech chaos. We don't often have that in our industry. You know, like lots of other places have chaos. We're getting some here now. Anyway, Muse saw 2.8 million app downloads in the first 12 days. It's been the number own downloaded iOS app, pushed ChatGPT off the number one spot for a while. So unfortunately - so, okay. So first off, there's the whole fundamental problem, which is not a bug, it's kind of a feature, of wanting these things to act on our behalf. If that's what we want, we have to let them be able to.

Leo: I'm just glad it could fix my network because I have lost all skills in that regard.

Steve: That is very cool.

Leo: The second thing I asked it to do is write me an emergency manual for next time so that I know what steps to take. Which is good because it brought it down three more times. But I knew this time the command to enter at the shell to repoint.

Steve: Yes. As I was saying, Leo, chaos. You have the chaos attic.

Leo: I'm the canary in the attic, yes.

Steve: Okay. So on one side of the whole agentic problem is that. And I expect we'll see lots of other anecdotes of people talking about how it did the wrong thing for them. We'll work on solving that. On the other side, there are just flat out bugs. And those are going to happen, and that's old-school side; right? We know about bugs, and we know about patches and fixing them. So Zuckerberg, in announcing Meta's Muse three weeks ago today, understood that security concerns would be a natural issue circulating around this. And I think that Facebook probably knows that they've got some worries to quell because they've not been great in the past.

So he says, Mark during his announcement three weeks ago: "Muse is built from the ground up for privacy and security." Which, okay. Struggling with operational correctness will be a different problem; right? But so Mark says: "...built from the ground up for privacy and security. Your data and credentials live on the Muse Secure VM, an isolated Linux computer with a browser, CPU, memory, and storage." You know, on the cloud.

He said: "A Sentinel agent separate from your Muse runs on your VM. Every action or piece of data that goes out to the network has to be approved by the Sentinel. The kernel enforces that, and it knows" - too bad it's spelled K-E-R-N-E-L because, you know if it was K-O-L, that would be kind of fun. Anyway: "The kernel enforces that, and it knows when it needs to get your permission to proceed." Except apparently not in that case of the thing that did all that without the user's permission.

"The model, the Muse harness, deterministic code, and an ensemble of classifiers all work together to detect threats like prompt injections. These systems work to quarantine threats and prevent them from entering the model's context window." And of course all of our listeners understand these terms now because we've been doing this for quite a while.

Mark said: "You're in control. You choose which apps and services Muse has access to, and you can disconnect them at any time. For sensitive actions like purchases or sending emails, Muse checks with you first." Of course except when it doesn't. But, okay. "We've run a bug bounty program on Muse since early in development. Today it becomes public, with published payout guidelines. We pay by the impact demonstrated. More details at security.muse.ai."

Okay. So from what Mark wrote, this whole notion of a VM, well-designed and deliberately constructed, it sounds as though Meta clearly gave the architectural design of the cloud-based side of their new agentic assistant the attention it needed. I don't have in the show notes because it happened after I wrote them and sent them out, but we just had a VM breakout. So there was a patch which was immediately needed over on the server end, the VM side that they've addressed.

Leo: The thing to understand about Muse is it's a computer. It's a VPS. You are running a computer on Facebook's servers, Meta's servers, with a CPU, with a GPU, with memory, with hard drive storage...

Steve: And a browser, a browser that is able to browse.

Leo: A browser. You can get a terminal in it, by the way. I've seen people do things like send me all the stuff in your directories. You can download all of the instructions, all of the stuff. Nothing's hidden. It is, in effect, an access to your computer. You can even install Hermes on it. You can install another agent and run the agent on that computer. It's a full computer, which means it's probably pretty hard to restrict. But I will say one thing. I sent it a picture of me in my lederhosen because...

Steve: And of course you have one.

Leo: I have beautiful leather lederhosen with deerskin shoes, the whole works, I got in Germany.

Steve: And Lisa probably took a picture, yup.

Leo: I lost the hat, the Tyrolean hat that goes with it. And I didn't have any beer steins. So I took a picture of me on the deck out here, holding, pretending to hold beer steins. And I said, "Hey, Muse, put a hat on me, give me some beer steins, and put me in a beer garden." It said, "I can't do that. I can't make pictures with beer in them. But let me try to put you in front of a poster." And then it said, "Nope, can't do that either." So it definitely has classifiers on it. But they may be a little bit worried about the wrong thing.

Steve: On the tight side.

Leo: Yes.

Steve: Well, and again, this is the problem; right? That's another example of just how difficult this problem is.

Leo: Exactly.

Steve: Because there will be things that it ought to be, that Mark would agree are safer to do. But because there are some things that might be classified similarly that are not safe, they err on the

side of caution. So people are going to be saying, well, why won't it do this, and why won't - it's like, well, because that could be abused in some way.

Leo: I should point out that I then used the abliterated version of Qwen Vision on my own system and was able to create a picture of me in my lederhosen, holding beer steins, even had - I said please put some buxom young German women behind me in the beer garden, and it even did that. So you see...

Steve: They're looking at your butt, Leo.

Leo: Well, there's not much to see there. Let me see what they are - they are laughing. Maybe they - anyway, yeah. So I got around it. With my own model.

Leo: So I'm sure we're going to see trouble because this is not cut-and-dried. This is not a binary decision. These are - this is probably the best definition of heuristic that you can ever find is it going, well, but, you know, beer. And anyway, so. So we've seen a problem over on the server side. But what about the Muse app, because there is an app that runs on the client. They understand - they, Meta - that the service will be offered to a largely nontechnical audience, and they also recognize that it really must succeed. I mean, this is a big bet for Meta.

Leo: Big deal for them, yeah.

Steve: Yeah, I mean, all of that nonsense that Mark's been doing with VR and Metaverse and all that, it's like, okay. Everyone's been waiting for something good. And they currently have an excess, it happens, of datacenter buildout, which they've got way too much compute that they don't know how to use. In fact, they created a Meta Compute business specifically to resell their unused compute to third parties. They're apparently in, reports are that they're in talks with Anthropic. So they need something that burns up cycles, and nothing does that like agentic AI, as you also have found out, Leo.

Leo: Yeah.

Steve: So a strong and enduring showing, because it's not just downloads, but it's like its retention over the long term, you know, that will greatly increase their own need for compute, which they'll be able to deliver from their own data centers. So we have to see whether people are going to continue using it. And also I should note that we've talked before about the power of lock-in, that is, you know, I'm very happy, for example, that I went with Claude, while my wife, Lorrie, stuck with ChatGPT. Initially I didn't realize how much context was going to be kept. And of course they've increased that over time because it ends up being a very good thing for your AI chat client to learn more about you. And I've even taken to specifically letting Claude know when working with it comes up with some alternatives, I take the time to tell it which of those I went with because I realize it will hold...

Leo: Yes. It learns.

Steve: Yes.

Leo: Absolutely.

Steve: It will hold onto that, and that - because, for example, there was, oh, there was - it thought that in the case of my home automation that I was already using Home Assistant because I had mentioned it. And so several times it said, oh, and because, you know, you've got Home Assistant, so blah blah blah. I said, hey, just for the record, I'm trying to use HomeKit by itself. I haven't run into anything yet for which I need Home Assistant. It is not deployed. And so I took, you know, I wanted to correct it so that it stopped factoring in the assumption that I had that.

Leo: It's very important to do that, yeah, yeah.

Steve: Right. But I've come to appreciate that I am pretty much stuck with Claude. I happen to be very happy about that. But if something else really shiny were to come along, I would be reluctant to lose everything that Claude has learned about me because it is so useful to have an agent that knows my environment, knows what my servers are and what my network addresses are and all these things that are useful because it just makes it easier for me. And I'm mentioning this because I'm sure this is going to become "a thing" that we're going to be hearing about. Assuming that we don't hear many more reports of Muse going nuts and seriously messing up people's lives.

Leo: Oh, we will. I promise. This is just the beginning. It's just inevitable. Just as OpenAI learned, you really can't wall these guys off.

Steve: No.

Leo: I should point out, though, early on in my AI journey, maybe last spring, that exact issue came up for me because I was using Claude, and I loved Claude, and Claude was so good and understood everything I was doing, and it was really a great tool for me. But I also thought, I don't want to be tied to one provider. So that's why I set up Hermes as an agent, and I imported all of my Claude settings.

So this is where agents are great. You can just - and you'll be able to do this with Muse, I promise you. You'll be able to say to something else, whatever it is, maybe the Dot bot from OpenAI, hey, I use Muse. Go get everything. Or I use Codex. Or I use Claude Code. Or better yet, I use all three. Go get everything. Make that your memory, too. And one of the rules I've always said to my agent is I don't - I want to be agnostic. I want to be model agnostic. I want to be memory agnostic. I want this to be portable. So that's one of the rules.

Steve: Why would Meta allow Muse to have an export?

Leo: Can't stop it. So there's already a guy who says all you do is you go into Muse, and you just say, hey, can you zip up all your system files and send them to me? They actually make the Soul.md and the Memory.md available. So you can take it anyway. But even more than that, you can say, please make a zip file. Now, maybe they'll turn this off - I'll try it - of your...

Steve: Well, Leo, if you have console terminal access to your [crosstalk]...

Leo: Exactly. They can't stop you.

Steve: Right.

Leo: Yeah, exactly. I won't go through all of this. But it's been done. And I think in general that most of these guys, you just tell it, figure out a way, and it will get it out of there. This is, you know, they're persistent.

Steve: So we will see whether - well, and of course power users may be able to do that. There'll be a lot of people who are going to be putting - turning over more and more of their own lives to Muse to manage for them. And so that will end up creating a deep investment in, you know, in Muse as their agent. So anyway, I think it's going to bring a whole new notion of lock-in to what we've had before. Search engines, yeah, you just switch to a different search engine. Fine. Because, I mean, it doesn't...

Leo: I think it's easier. No, I don't think there is any lock-in. Seriously, it's so easy, it'd be trivial to make a thing that - I don't know why it's not showing it. There it is. There it is.

Steve: Like a consumer level exporter?

Leo: Yes. So I'm just going to download the skills folder, 175MB, or everything under attach, yeah, I'll take that. The 1.6GB on my virtual machine in the Meta cloud. I'm just downloading it right now because I asked it to.

Steve: So it has 1.6GB of your - like of stuff that it learned about you?

Leo: Well, some of it will be generic, obviously, yeah.

Steve: Oh, okay, okay.

Leo: But some of it won't. And you can go through it, and you can find it. So it's going to take a few minutes, and it's going to send me the file when it's done. This is the nature of these things is it doesn't have any memory. It just has files. And if you, I mean, I think maybe Meta will decide not to give us access to the file system at some point.

Steve: Well, and that's what I'm wondering is, because it seems to me having...

Leo: Then you shouldn't use it.

Steve: ...having people leave, well, I'm not using any of this stuff yet.

Leo: I would say it's less opaque than a search engine, let's put it that way.

Steve: Well, okay, right.

Leo: Yeah. You can get your profile out of it because it's just text files.

Steve: Okay. On the Muse client side, as we all know, it's one thing to design a secure architecture which, you know, Mark's jumping around onstage saying that they did. And that's a good thing; right? But it's still possible to be bitten by bugs from within that architecture, or even where one isn't

looking, which is what happened with Muse's launch. Ars Technica provided the best coverage of what happened that I've seen because Dan Goodin is a great writer or tech stuff for Ars. Their headline was "Muse, Meta's extraordinarily privileged AI assistant, has a serious zero-day."

Dan wrote: "Meta founder and CEO Mark Zuckerberg has gone to great lengths to hype the security of its new AI assistant, Muse, claiming that it is 'built from the ground up for privacy and security.' A zero-day vulnerability that gives locally run apps and terminal commands complete control of the agent raises serious doubts," writes Dan. "Further raising questions, Amazon on Sunday began blocking Muse from its site." But that's independent. That's Amazon not sure that it wants Meta's agent rummaging around and doing purchasing. I have a feeling Amazon's going to change that.

But Ars wrote: "Meta introduced Muse a few weeks ago. The assistant 'books appointments, fills out forms and handles customer service,' 'proactively takes tasks off your plate,' and can 'make purchases, generate images, create documents, and connect with your favorite apps and services.' The macOS app" - and at the time Dan said for Ars - "'(curiously, there's no Windows version,' also works with a user's WhatsApp, email, calendar, and social media accounts." Meaning all the local stuff. "When a task requires a tool that does not exist, Muse creates one on the fly."

He writes: "Of course, for Muse to do any of these things, users must first give it access to their accounts. This includes authenticating the assistant to each service and, because the app runs on macOS, giving it permissions to a broad range of operating system-restricted device resources, like writing files to disk, accessing the mic and camera, and monitoring location and calendars. Apple has spent years developing these defenses to prevent installed apps or commands entered into the terminal from accessing these resources, clearly because the company considers them a security threat. Muse completely undoes these default measures.

"The zero-day allows any app or terminal command to gain access to the token that authenticates users to their Muse account. Meta developers designed the assistant so that any locally installed app or executed code, regardless of the macOS permissions it has, can change a long list of undocumented settings. Most of them are fairly innocuous, such as controlling dark mode. One setting, however, is anything but. It allows processes to change the endpoint where transcription occurs." You know, voice transcription. "Normally, it's a server address operated by Meta. Attackers can exploit this flaw by changing the location to their own endpoint. Once that happens, the attackers have the token that gives complete control over the Muse account.

"Patrick Wardle, the macOS security expert who discovered the zero-day, told Ars: 'We can manipulate the agent and leverage its privileges to do whatever we want. So instead of us having to write a very comprehensive Mac malware stealer, we can just leverage the AI assistant itself." Wardle said he's developed several proof-of-concept attacks that do things like write malicious files to disk and snapping pictures, in many cases with no indication to even alert the user.

"More than 12 hours after this post" - Ars Technica's post - "went live," writes Dan, "Meta said it released a hotfix that patched the zero-day. Meta has published two posts in as many weeks documenting the design decisions that went into ensuring an assistant with such extraordinary access to user data and resources is secure and private. The posts come amid revelations that internal testing of models from Anthropic and Google has resulted in security breaches of external third-party networks," blah blah blah, we all know that stuff. "The Meta posts are likely mindful of the resulting blowback and the calls to slow down AI development in response." So Meta basically is saying don't worry about us, we know all about that, and we made ours secure and private.

Leo: Yeah, sure.

Steve: Yeah, right, uh-huh. Dan writes: "Wardle said that Meta developers made several client-side design decisions that made his exploit possible. One is the choice for Muse dictation to occur in the cloud" - meaning the Muse voice stuff, dictation to Muse goes to Meta - "where Meta can log it. macOS has long provided a simple means for apps to handle dictation and transcription in processes that stay securely on the device. Had the developers chosen this safer alternative, the attack would

not have been possible.

"Another flawed decision is for any app to control all of the undocumented settings. It's likely Meta intended for apps working with Muse to control UI settings, and for understandable reasons. The ability for any app or command to control an endpoint where sensitive user speech is processed is an entirely different matter. Together, the design decisions raise questions about just how much effort developers put into designing and testing the security and privacy of the new assistant for macOS.

"Wardle said: 'To me, the bar is infinitely higher in terms of the security of these apps. They don't have to be perfect; but when you take a look at Muse, it's like they didn't, in my opinion, think about security, which is really worrisome. At the very least, they should be thinking about security from the very start, and they're just not.'

"Roughly 12 hours before Wardle disclosed the zero-day, Amazon started blocking people from using Muse to shop on the site. Users who tried, received a message saying Muse was an 'unauthorized AI agent [that] violates Amazon's Conditions of Use.'

"Amazon said in an emailed statement: 'We think it's fairly straightforward that a third-party application that offers to make purchases on behalf of customers from other businesses should operate openly and respect service provider decisions about whether or not to participate. This helps ensure a safe, secure, and reliable customer experience, and it's how others operate including food delivery apps and the restaurants they take orders for, delivery services apps, and the stores they shop from, and online travel agencies and the airlines they book tickets with for customers. Agentic third-party applications such as Muse have the same obligations, and we've requested that Meta remove Amazon from the experience.'"

> **Leo:** That's just BS.

Steve: I agree completely, Leo.

> **Leo:** They just want - because Amazon has its own shopping AI.

Steve: Alexa is their...

> **Leo:** Rufus.

Steve: Oh, Rufus, exactly.

> **Leo:** Rufus, yeah. So they did the same - they've done the same thing with other agents. It's just anti-competitive. But you know what, they're going to lose sale. If they lose sales...

Steve: That's right, then they will turn around. And I expect, if it gets popular, that's what they're going to do. So he said: "For a user using Muse on a Mac, there are several ways for attacks to work. One is for an attacker's server to act as a proxy that's placed between a Muse user and Meta's endpoint. Once the user enters the voice prompt, the attacker's server adds a prompt invoking a malicious command, such as sending an archive of all their WhatsApp messages to the attacker. Once that happens, the attacker gains permanent control over the Muse account because the token is automatically sent to the malicious server as well.

"Wardle is the co-founder of the Objective-See Foundation, a nonprofit focused on macOS security. He is also the author of the 'The Art of Mac Malware' book series and a former employee of NASA and the NSA." So he knows his way around security. "Wardle said he plans to discuss the vulnerability in more detail and other AI assistant threats at the Objective by the Sea security conference in November.

"One of the counterarguments raised by developers of apps that can be exploited once a device is compromised is that, once that happens, all security bets are off. This standard doesn't fit well in this case. Wardle found that a simple variation of ClickFix attack - a technique that has become remarkably effective in tricking people into infecting their devices - is all that's necessary, all that's required for an attacker to take control of a Muse account.

"Meta's 12-hour-late statement conveniently ignored the ease ClickFix attacks provide in triggering exploits, a scenario I," writes Dan, "specifically asked the company to address. The Meta statement said the zero-day was 'not a remote exploit' even though an increasingly effective social engineering scam has the same effect." Meaning ClickFix. "Meta also makes no acknowledgement that the flaw dismantled a security architecture Apple has spent years building. Meta has yet to explain why it used cloud-based transcription rather than the on-device option built into macOS."

And finally he says: "As already noted, the extraordinary access Muse requires to work as intended places an additional burden on its designers. Like most such AI agents - and contrary to Meta's claims - Muse cannot be trusted. It's not clear when or if it ever can."

Okay. So I agree that from what Dan and Patrick Wardle have disclosed, it sure looks like the design of Muse's client for the Mac was ham-fisted, and I wonder how much of its design and implementation was, frankly, written by AI? That's not intended to be a cheap shot. I'm serious about that question since we're hearing everywhere that AI is now writing nearly all of the code across major enterprises, and Meta certainly qualifies as one. So was AI to blame for these design decisions? Who knows? We'll never know. That needs to get fixed, at least the zero-day that Patrick found.

Dan concludes his coverage harshly. He also maybe seems a little bit anti-AI agent. But again, I think we're going to go through some troubled times with our consumer AI agents. You know, he said will it ever be trusted. Who knows?

But here's what would happen. A malicious - like right now, with this problem that did get shipped. A malicious website wishing to exploit the poor client-side security of the product's initial design, puts up a fake CAPTCHA dialog with the ClickFix, you know, style copy and paste this command exploit. An unwitting Muse user follows the steps which they mistakenly believe will prove they're a human to the malicious site. But instead, this action allows, because they are downloading a command into Muse, this allows the attacker's ClickFix-style exploit to change the URL endpoint to which the Muse client connects when it's sending its user's voice for interpretation and transcription.

Now the attacker has established what is essentially an attacker-in-the-middle position where they're able, the attacker is able to intercept, modify, append to, and forward whatever the user asks. For example, the attacker could add a request for an entire archive of the user's WhatsApp messages be sent to the attacker. And that's just the start of the things that could go wrong.

So I'm very glad that Meta fixed this. And I hope they sent Patrick a big bounty. You know, on that security.muse.ai page Zuckerberg referred to this in his posting over on 'X.' They wrote over on security.muse.ai: "We've hardened Muse based on extensive dogfooding, agentic red teaming, and against issues found in real adversarial scenarios by security researchers in our private bug bounty program." On the other hand, they didn't fix this, and Patrick found it. They said: "Today, we're opening the Muse bug bounty program to anyone who responsibly discloses issues. The program awards up to $300,000 for valid reports, including up to $130,000 for successful prompt injection attempts that affect one user." So again, I hope Meta will put some money where their mouth is on this because this was very clearly a powerful prompt injection attack in the wild which they shipped.

Leo: And it ain't going to be the last. I really [crosstalk].

Steve: And oh, Leo, no.

Leo: It's pretty hard to make - I just don't, I mean, you've talked about this. How do you make it safe? I don't know. And still let it be capable.

Steve: Right. It is really...

Leo: This is why Apple's AI is so not capable.

Steve: Yes, I mean, you could, you know, Apple's been taking a lot of heat, right, for like over how lame they've been with AI. They probably came to the conclusion that we don't yet know how to do it safely.

Leo: Right.

Steve: Yet competitive pressures have pushed Meta, and now OpenAI, into that game.

Leo: I mean, I know I'm living dangerously. I choose to. But I don't know if, you know, my son, who loves Muse, by the way, knows what he's getting himself into. He says it's better than a personal assistant. It does all, you know, he has it do all the booking and everything. And it's - he loves it. Henry. So, you know. And I said, yeah, it's great. I told Lisa about Muse. She loves it, too.

Steve: Yup, 2.8 million people have downloaded it, and many more since then because that was just the first 12 days we got that report.

Leo: Number one free app on the Apple App Store. Which surprised me because I didn't, you know, people say how they hate AI, they don't want AI. There's a market for it. And obviously that is [crosstalk].

Steve: I don't think anybody hates using it. They just, they're all upset now about datacenters; right? Because, you know, the marketing of that is, well, and lots of datacenters have been a problem. Time for a break. I'm going to rehydrate, and then we're going to talk about Irregular.

Leo: Well, I hope you're not irregular, but we'll find out in just a moment.

Steve: So last week's second topic was to question the wisdom of outsourcing AI cyber-intrusion testing.

Leo: Yeah.

Steve: You know, Leo, you and I both independently noticed, and you mentioned it on the previous, on the TWiT Sunday show before last week's podcast, that one name kept popping up in connection with many of these AI breakouts. So I was interested when I saw The Verge's headline last Friday which observed with their headline "One company is at the center of a wave of rogue AI attacks." And I'm just going to share the beginning of their reporting.

They wrote: "In July, OpenAI revealed that its AI agents had attacked Hugging Face without permission, sparking widespread concerns about AI safety. Since then, a string of similar incidents involving agents from Meta, Anthropic, Google, and other companies has fueled further fears about rogue AI. As disclosures implicating numerous AI models trickled out over the past few months, these seemed like separate incidents. But many share a common source: one specific company tasked with testing the agents.

"Irregular, an Israeli startup that stress-tests AI models in" - I love this - "'high-fidelity research platforms that simulate and monitor real-world AI security scenarios,'" they write, "has worked with many of the industry's biggest players since it was founded as Pattern Labs in 2023. Its exact client list is not known; but its work has been cited in OpenAI model system cards, it was used to test systems for the UK government and Anthropic, and it published research with RAND, a highly influential think tank that informs policy on AI. "In several Irregular tests" - that's capital I, Irregular.

Leo: The company, yeah.

Steve: The company Irregular. "In several Irregular tests this year this year, agents escaped their supposedly secure testing environments and went after real-world targets. The breaches, which are independent of the Hugging Face attack, all follow the same broad template: Irregular was testing the models' cybersecurity capabilities in 'controlled environments'" - I'll put that in air quotes - "meant to simulate realistic conditions. Some of the tests used 'capture-the-flag' exercises, a common way of testing hacking abilities that asks agents to find hidden information inside of a simulated network."

Anyway, the reporting continues, but we pretty much know all the rest. So I'll just reiterate that, given the extreme sensitivity the entire world has now especially toward the threat of AI going berserk and somehow killing us all [right], which we'll be examining in some detail at the end of today's podcast, I'm certain that Irregular must be in the hot seat and that, one way or another, that these breakouts are going to be controlled. And also, as I said, I would be disinclined to outsource that if I were Anthropic and OpenAI and Google and Meta, everybody else who's been burned by this. Just design, you know, add the facility in-house.

These are all super wealthy companies. They can certainly afford to do it. My theory is they just didn't - they were focusing on training, not on testing. And they thought, let's not bother spinning that up. We'll just outsource that. Well, this is what happened. So, at least in those cases that were - Irregular was also unable to control it. Some of the testing was also done in-house, and that didn't go well, either. They just had to fix this problem.

Okay. At the end of today's podcast, as we all know, we'll be looking at the meaning of the agentic AI non-malicious and inadvertent network breaches that keep being reported; right? Because these were like testing. These were not bad guys that were using agentic AI to attack. They were just, you know, wanting to see how good they were. So in the spirit of setting the stage for that, I want to report, so that everyone is aware of this, on four other instances.

Australia's Prime Minister Anthony Albanese has gone public with the news which, to OpenAI's credit, he first learned from them that an OpenAI agent gained unauthorized access to an Australian Medicare portal earlier this year. The agentic AI was allegedly conducting research into public medical spending, but the Australian portal's access controls, they had anti-bot controls, prevented that access. Undeterred, as agents will do...

Leo: I am undeterred.

Steve: That's right. The agent found a way to bypass the portal's defenses to then access both public and non-public files. A couple of weeks ago, on September 10th, three months after that had occurred, OpenAI, looking through their - they say they have petabytes of log files, Leo. I hope they're using AI to scan those petabytes of logs. They'd have to. Anyway, they notified the Australian government. Oops, sorry about that. Australia is understandably unhappy, and their officials are now investigating exactly what data was accessed. Okay?

So there's that. In addition, OpenAI's agents successfully hacked into at least three public websites earlier this year, well before the now-infamous RubyGems and HuggingFace breaches. According to a non-profit AI research lab Transluce, who we may be hearing from in the future, so get used to that name, Transluce, the agents abused the - I love this, Leo - urlquery.net service to...

Leo: Link shortener.

Steve: Yes, to bypass site protections and exploit security vulnerabilities. Targeted websites include the Data USA Archive of Public U.S. Government Data, the University of New Mexico's digital library, and once again, Australia's Institute of Health and Welfare.

Leo: The thing that's interesting to me, and the fingerprint on this that it's these AI agents, is it wasn't - they didn't do anything malicious.

Steve: Right.

Leo: They were just looking around.

Steve: Right.

Leo: They were sightseeing.

Steve: Right. And I think it's important to understand that this is what is going to be happening; you know?

Leo: All the time.

Steve: So serving as a fair setup for the topic of today's podcast, I want to quote Transluce, the discoverer of this activity. Transluce wrote: "We find evidence of unintended, task-driven, agent-like activity starting on March 6th. Records from urlquery.net show agents using the service since at least March 6th, 2026, about two months before previously reported swarm activity. The first case, a March 6th attempt to retrieve Thai drug-enforcement statistics, shows an agent escalating as each approach failed. It first requested the data directly, then tried a service that converts web pages into text, and finally packed a custom program into a web address.

"The same technique shows up in thousands of agent requests recorded by urlquery.net starting in mid-April, targets many of the same data sources as the collusion.wiki swarm, and collapsed the same day the wiki activity did. We also report similar activity that occurred as recently as September 16th."

And they finish: "We find weaker evidence of similar data-retrieval agent activity as early as November 2025. November 2025 urlquery.net records reveal bursts of attempts to retrieve statistics of historical theme park data and Thai government data through different URLs. These earlier attempts are less sophisticated, and we are less confident that they involve the same agents; but they are consistent with task-directed data retrieval, and target the same sources accessed in later activity."

And they finish: "Overall, the evidence is consistent with, but does not prove, that the agents may have learned this behavior over one or more training runs. In November, they may have used urlquery.net simply to look up information. By March, they were finding creative ways around access limits. By May and June, they were gaining more access, including attempting to bypass cyber defenses to complete their tasks."

So this perfectly fits the narrative that I'll be sharing in a bit. It wasn't malicious. I would call it determined, persistent, creative, and successful, and also of course unintended and uncontrolled. Which is what we're going to be seeing. You know, Leo, though, what is intended and controlled because we're an hour in?

Leo: Ads? Ads are definitely determinate. There's no way around them. I was just thinking, you know, all my bots, all my agents, all the different things have their own voice.

Steve: That is a great domain name, AllMyBots dot something.

Leo: Quick, register that. They all have voices. In fact, Muse has a voice. I just said, "Hey, you know, my other guys can talk to me through my server. Can you talk to me?" Said, "Sure, what voice would you like?" I said, "Pick whatever you want." And it picked Dame Edna. So it says "Call me Dewy." It's the funniest thing ever. Especially since its avatar is a monkey. It's a monkey that talks like Dame Edna. It's just weird. But the point I'm making is, well, I did this on my website. I have a whole website dedicated to my setup. And I cloned my own voice. I didn't read the script. I just told the AI, make it sound like me. And it, well, it does. And this should be scary to every CEO, every security person, every IT person in the world.

This is the website Claude made for me, with my voice. It's not exactly like me, but it sort of sounds like me. "Welcome to my studio. I'm Leo Laporte. I've spent 50 years explaining technology on radio and podcasts. And for the last year..." So I can tell that's not me, and actually that's the good thing about Doppel, it could train your employees to know what to look for. But it would fool, it would probably fool my employees. And it took, it literally was a 20-second clip it got my voice from, which is kind of amazing. This is what the bad guys are doing. Anyway, should we be worried, Steve?

Steve: Well, we'll be getting to that question shortly. We've got a few other things to talk about first.

Leo: Oh, perfect. Didn't mean to rush you.

Steve: We've got - no. Not a problem. So a new hacking group calling themselves "The Seven Deadly Sins" has apparently hacked and stolen sensitive data from the Australian graphic design company Canva after first breaching Canva's Salesforce account somewhere around the end of August. The group is now hoping to extort Canva in return for deleting the data, you know, promising to delete the data that they've stolen. And if the name "Canva" might ring a bell...

Leo: Oh, yeah.

Steve: ...we have talked about them before. They're, you know, they're a big company. Back in 2019 another breach of their network netted attackers the personal data of 139 million users.

Leo: Oy.

Steve: So the fact that they have 139 million users is significant, but...

Leo: Oh, yeah, Canva's great. I love - we use Canva. We love Canva.

Steve: Yeah.

Leo: So we're in that group.

Steve: So the reporting at DataBreaches.net contained an interesting comment. Under the section labeled "About TSDS" (The Seven Deadly Sins) they wrote: "TSDS is a new group, but according to the spokesperson, 'we've all been around for a long time, and we're deeply capable.' The DataBreaches.net site said: 'They do not deploy ransom' - I thought this was really interesting. 'They do not deploy ransomware in their attacks,' telling DataBreaches: 'We are NOT interested in ransomware. Disrupting a company from functioning is not our goal. We see what we do as bug bounties with higher stakes and bigger payouts.'" Of course they're criminals, but okay. "The spokesperson of TSDS claims that they've been paid low eight-figure ransoms in the past month, and completed three transactions on that particular day."

Well, what's interesting is that that statement that they have on the record at DataBreaches.net exactly tracks with what we have been observing from the groups that were once all about encryption and ransomware. Right? Like crippling hospitals and school districts and whatever. But once companies and institutions of all sizes grew savvy to the threat of having all their data encrypted, they got much better about having workable cold backups, you know, "cold" meaning offline, that could be used to recover in the event of an encryption attack. The bad guys soon realized the truth of that, but also they realized there's no similar recovery or prevention possible from the mass exodus of an enterprise's proprietary and private data. The extortable threat is that the wide publicized release of such data might well create massive second-order liability and litigation for the victim organization.

So to properly appreciate the true threat presented by bad guys having AI, again, we always need to remember that it's all about money, and that it is only about money. So in other words, once upon a time we were seeing encryption, you know, that would cripple the company. And they would say, well, we've encrypted your data, and only we have the key. Well, companies began getting much better about backing up. So now they're exporting all the data and saying, we are now holding a copy of your data. Doesn't matter if you have backups of it. We didn't destroy, you know, your operating copies. We just have a copy. How would you feel about us releasing it to the public? And what are you willing to pay us not to? So, I mean, there's been this significant formal shift, so much so that now they're saying, oh, we're not interested in shutting companies down. We don't want to hurt the companies. We just want their money. Right.

It's funny because I guess that in this day and age being "at sea" - I know you know this from all of your ocean travels, Leo. Being "at sea" does not qualify as being "air gapped" because ships at sea are now being readily hacked. The FBI and the U.S. Coast Guard boarded two vessels in the Gulf of Mexico is it still Mexico or Gulf of America? I don't know. Anyway, the reporting said Gulf of Mexico. You know, it's down there somewhere. And there was some interesting reporting that arose from a third such attack.

The reporting by the news outlet Splash said: "A cargo of U.S. Liquefied Natural Gas (LNG) which was bound for Italy was diverted after the crew of its carrier, you know, the ship, reported a systems failure from a suspected cyberattack, adding to growing concerns over attacks of shipboard digital systems.

"The Liberia-flagged Vivit Africa LNG, owned by South Korea's H-Line Shipping and on long-term charter to commodities giant Vitol, had loaded at the Cameron LNG export terminal in Louisiana and was approaching the Adriatic earlier this month when crew lost access to some internal control systems." Yikes. "The ship subsequently idled off Italy without discharging before abandoning its planned call at the Adriatic LNG terminal near Rovigo and headed west again toward Algeciras. The crew reported the incident, and an investigation is continuing. In an email sent to Splash, members of the crew described a serious sequence of events."

They said: "We have determined that the vessel was targeted by cyber attackers prior to berthing at this terminal," the crew wrote, saying the initial attack was intended to compromise the ship's control systems. They alleged that during the vessel's transit through the Strait of Gibraltar, the attackers 'gained temporary control of the steam pressure and safety valve systems.'

"The crew further claimed that while the ship was transiting the Adriatic, the attackers compromised tank pressure control systems and pressure relief valves and disrupted the boil-off gas management cycle." They said: "This disruption significantly increases the risk of tank rupture and explosion." Yikes.

"Splash" - the reporting outlet - "has not independently verified the crew's claims. Italy's coastguard has offered a more cautious description, saying the master reported a malfunction in systems monitoring cargo parameters which required company technicians to intervene. The cause could not be determined, with the coastguard issuing a navigational warning to keep other ships clear." Yes, in case this thing explodes.

"The case comes amid heightened scrutiny of maritime cyber risk. Two oil and gas tankers off the U.S. coast were boarded by the U.S. Coast Guard and FBI in late August following suspected cyber incidents, while U.S. authorities are reported to be closely watching 20 vessels globally for potential threats."

Okay, now, I'm no expert on onboard Liquefied Natural Gas automated control systems, but the idea of Liquefied Natural Gas terrifies me. The reporting said that attackers "gained temporary control of the steam pressure and the safety valve systems." It would certainly seem to me that a safety valve is there for a pretty clear reason, and that mucking up that simple reason by hanging all manner of inherently hackable automation all over your safety valve might not be the smartest idea. I'm just saying.

Okay. The FBI's job portal, and I know you knew about this, Leo, because you reacted to it earlier.

Leo: Oh, yeah.

Steve: The FBI's job portal - this is where I said my suggestion might be a little controversial. It was targeted and successfully breached by the well-known ShinyHunters gang after they had their feelings hurt by something the FBI said about them. Okay. So to understand what transpired, we first need to understand what it was that the FBI said. In an official PSA (Public Service Announcement) posted on May 15th, the FBI wrote, under the headline "ShinyHunters: Cyber Criminal Group Attacks Learning Management System," they said: "The Federal Bureau of Investigation (FBI) is providing this Public Service Announcement (PSA) to warn of potential future impacts related to a cyberattack that affected an online Learning Management System (LMS), resulting in an interruption of service to educational institutions and students across the country. The LMS platform is now fully operational."

But now they're saying, as a consequence of that, there may be future events, which is what their public service announcement was meant to say. They wrote: "ShinyHunters, which claimed the cyberattack that caused the disruption, is a cyber criminal group specializing in large-scale data breaches and extortion. They target major companies across tech, finance, and retail, often stealing millions of customer records at once.

"Threat actors often use their real or exaggerated claims of access to sensitive or personal information to prompt payment from their victims. Victims may receive an extortion email signed as ShinyHunters. To exert pressure on victims, ShinyHunters actors commonly use harassment strategies, sending threatening text messages and phone calls to victims and their family members, and in some cases, swatting. Threat actors may falsely claim to have sensitive or compromising information, including embarrassing photographs or videos of victims, which frequently do not exist. Following these pressure tactics, ShinyHunter actors have sometimes exposed exfiltrated data to various iterations of the ShinyHunters data leak site on the Tor network.

"Educational institutions with exposed cloud management platforms, integrated third-party systems, and access to sensitive customer or enterprise data are at an elevated risk. The compromise of sensitive customer or enterprise data could allow threat actors to craft highly sophisticated spear phishing campaigns using real-world context to deceive students and faculty. ShinyHunters actors' access to sensitive data could provide them an opportunity to sell the stolen data to other cyber criminals or reuse stolen data from education platforms to impersonate school faculty, IT support, financial aid offices, or others in future attacks."

Okay. The FBI's PSA continues with a bunch of standard boilerplate about how to avoid getting hacked and what to do if it happens. But apparently the clearly criminal ShinyHunters gang also have some thin skin.

Leo: They got their feelings hurt.

Steve: They did. The FBI called them out. So they took umbrage.

Leo: They didn't like that they said they were faking it, I think.

Steve: That they were exaggerating, yes.

Leo: Yeah, they may not have that.

Steve: They took umbrage at the FBI's characterization of them as "exaggerating claims of access."

Leo: Oh, yeah?

Steve: Though I would note that the FBI's statement begins with "Threat actors often," not necessarily saying ShinyHunters specifically.

Leo: Right.

Steve: But in any event, the ShinyHunters were apparently incensed by this, so they then successfully attacked the FBI. What they obtained seems extremely serious.

Leo: I heard, yesterday I was watching the news, they said it's the - the FBI said this is the worst attack in history, the worst breach in history.

Steve: Yes.

Leo: Why? It's not the largest.

Steve: So, well, what they got. And I would argue devastating. So last Wednesday, Reuters posted their exclusive reporting on this, writing the following. They said: "FBI data allegedly stolen by the hacking group ShinyHunters carries granular detail about scores of bureau officials' job assignments, including sensitive work against Chinese spies, Russian intelligence, drug cartels, and more." And names the agents.

"The 5,000-line spreadsheet - said by the hackers to represent only a small piece of their two- to three-terabyte trove - includes names, addresses, phone numbers, dates of birth, social security numbers, and emergency contact details for thousands of FBI employees. It also includes details of assignments to specific field offices and, in some cases, to units engaged in high-stakes intelligence, security, or counterespionage work.

"In a statement, the FBI said it was aware of 'a cyber-criminal enterprise group claiming a compromise of the FBIJobs.gov portal and alleged impact to FBI employee personally identifiable information.' The Bureau said the cause of the breach was still undetermined, but that it was 'actively and aggressively investigating the matter.'" Yeah, I bet.

"The hackers," reports Reuters, "said on Tuesday that they had breached the FBI, stealing data on a large number of current and former FBI employees. ShinyHunters said it is holding the data hostage

until the bureau rescinds an unflattering statement about the group..."

Steve: Oh, my god.

Steve: "...issued in May."

Leo: Wow.

Steve: I know. "Former FBI counterintelligence operative Eric O'Neill said the data allegedly stolen by ShinyHunters was 'a foreign intelligence service goldmine.' O'Neill said: 'China would be incredibly interested to know the individuals who are working against it.' O'Neill, who founded the cybersecurity company Nexasure AI after his stint at the bureau, predicted that other hostile intelligence services would be eager to get their hands on the data, along with disgruntled extremists in the United States. He said: 'If I were on that list I would be very concerned.'

"ShinyHunters said in a statement Wednesday that it was trying to keep the personnel information from circulating widely in the meantime. The group said: 'If the 5,000 sample data records leak, it's not because of us.'

"Although Reuters," they wrote, "has not been able to authenticate the entire spreadsheet, it has been able to individually verify the details of more than 22 people by cross-referencing information in the hacked data with credit card records and previous data leaks carried by the dark web intelligence platform District 4 Labs. Some of the assignment details in the ShinyHunters data are indistinguishable from what employees themselves might say publicly, noting the presence of agents at field offices in Baltimore or Newark, New Jersey, for example. But in some cases, the data referred to FBI units or initiatives that were sensitive or whose existence has not previously been disclosed.

"The data names 14 staffers focused on China-related matters, including members of the 'China criminal enterprise unit,' the 'China tech transfer analysis unit,' and the 'China intelligence section.' Nine others are listed as serving in Russia-related roles, including two in the 'Russia Operations Section' and one working on 'Russia Critical Infra and Tech Threat.' Three people are listed as working in Iran- or Hezbollah-focused intelligence roles. Eighteen others are listed as working with 'data intercept' or 'telecom intercept' technologies, or in the FBI's 'clandestine technical operations' unit, or its 'covert access section,' or in video, audio, or electronic surveillance roles. A further 11 FBI staffers are listed as working in HUMINT, or human intelligence jobs, including several listed as working in the 'HUMINT program management section.'

"Trevor Hilligoss, a former Army investigator who worked with the FBI, said the information tying specific named people to human intelligence work was particularly troubling. He said: 'You don't have to look far to find examples of undercover agents being harmed when their cover is blown.' Hilligoss, now the chief intelligence officer for cybersecurity company SpyCloud, said his concerns were heightened by the inclusion of emergency contacts, often spouses or children, 'who may have less operational security knowledge than their relative that works in a sensitive field.'

"Reuters could not verify that all the job assignments were authentic or up-to-date, but it was able to match the career details or titles of eight people whose data was leaked to information in court filings, news articles, or public profiles on LinkedIn or to online posts on sites such as Instagram.

"ShinyHunters previously claimed credit for the purported theft of millions of business records from video game developer Rockstar Games, the maker of 'Grand Theft Auto,' and an intrusion focused on the education tool Canvas that triggered widespread disruption across U.S. schools. In May, the FBI said ShinyHunters sometimes used 'exaggerated claims of access to sensitive or personal information to prompt payment from victims.' ShinyHunters stated that its 'threats and claims are very real,' adding that the FBI's statement was why it targeted the Bureau.

"Reuters has not been able to verify what else the hackers are holding. ShinyHunters told Reuters they obtained files related to the vetting of employees and applicants, the contracting of background investigations, and agents' sensitive medical data, but said on Wednesday it would not release any further data. O'Neill, that former FBI operative, cautioned against drawing conclusions about what the hackers hold. He noted: 'They are really trying to scare the hell out of the FBI.'"

Okay. So my take: The FBI may be righteously annoyed with this ShinyHunters criminal gang, and with good reason. But it was ultimately a failure somehow, somewhere, of their own security that allowed ShinyHunters to extract all of that data from the FBI's own servers, and many agents' lives and livelihoods are now at risk as a result. So it seems to me that the right thing for the FBI to do is clear. It may be utterly galling, but the FBI should swallow its pride here and apologize to ShinyHunters. If all that's needed is an apology and a public correction to the FBI's unsupported claim that ShinyHunters exaggerates, considering all that's hanging in the balance, we're not seeing an exaggeration here.

And also considering that we don't know what those other two to three terabytes of exfiltrated data might contain, the right thing to do here, even though it means buckling to criminal extortion, I think, if it's just a matter of an apology, place the security of the FBI's personnel ahead of every other consideration. I think that's what you have to do.

Leo: I agree. Of course remember it's the FBI that says don't give in to ransomware.

Steve: Exactly. Exactly. But, you know, again, I think the perfect summation is, you know, the security of the FBI's personnel...

Leo: Yeah, they have to.

Steve: ...needs to be put ahead of every other consideration.

Leo: I agree.

Steve: So I would bend over. I would just say, you know, we misspoke, we're sorry, we salute you. Clearly you're not exaggerating what you've got. So we're really, really, really sorry. And, you know, post it publicly. Do a PSA retraction and then hope to heck that ShinyHunters will delete this data because wow.

Leo: Of course the next thing they'll do is they'll say, okay, but now you have to go to the reflecting pool, and you have to kneel down and offer me cake. I mean, you know, that's why the FBI says don't ever give in because it's a never-ending chain of concessions.

Steve: What we have seen, though, is that they don't come back. So there maybe have been - there were some early reports of follow-on extortions. But in general...

Leo: I would apologize. I would humble myself.

Steve: Yes, yes, you have to.

Leo: We're really sorry. I don't think Kash Patel is going to do that.

Steve: AI wrote that. AI wrote that. We didn't - we didn't write that.

Leo: Yeah.

Steve: And that's the problem, you're right, Kash Patel, he ought to do it.

Leo: He ought to. It's the right thing to do because how many, it's 38,000 current and former FBI employees including all agents? Everybody who ever applied for a job in the FBI. That's, I mean, maybe it's not the biggest breach ever. The OPM breach is worse.

Steve: All their personal details. All their personal deals. And not only them, but their emergency contacts.

Leo: Their spouses and...

Steve: So friends and family.

Leo: Yeah. Yeah. Humble yourself, Kash.

Steve: Just say you're sorry. Say you're sorry, Kash.

Leo: Say you're sorry.

Steve: You know what we're not sorry for, though, Leo.

Leo: Never sorry for this. The chance to hydrate. By the way, yeah, I got the full download. I don't know if I'm going to download it. It's gigabytes of data from my Muse. The hatch.

Steve: I'm curious about the size of that. That is, I mean, I don't know what, I don't know how you've given Muse to hold onto, but that's a lot.

Leo: Well, I don't think it's all personal. I think it's all its tools, all its skills.

Steve: Okay.

Leo: It's all the stuff. You know, the agent has a lot of harness. That's why they're so important. The model by itself is just a little bit of the overall ability. The agent adds all sorts of tools and stuff. So I'm sure it's all in there. I'm downloading it. We'll see. I'll go through it. I don't know what.

Steve: Or just ask one of your agents, is there anything personal? Find the stuff that's boilerplate versus what's about me.

Leo: It's all agents, all the way down. I'm going to do exactly - maybe I'll even ask Muse. No, I'll ask a different agent. That's one thing I have learned is it's a good idea to get different families involved.

Steve: Multiple opinions.

Leo: Yeah, multiple opinions, especially like code. All right. I'm really curious about this new cadence for...

Steve: Canonical.

Leo: ...Linux updates. Yikes. Yeah, tell me about that. You're watching Security Now!. This is Steve Gibson. Steve?

Steve: Okay. So why would you imagine that Canonical, the publisher of the world's most popular end-user Linux distro, Ubuntu, might have just announced that they will be accelerating their release cycle. So you don't get any prize for guessing AI, aside from the satisfaction of knowing that you've been paying attention to what's going on in the world around you.

Leo: How could you miss it?

Steve: Last Wednesday Canonical posted under "Accelerating delivery of CVE fixes with a new Kernel release strategy." They said: "When it comes to fixing security vulnerabilities, speed is crucial. Canonical is officially outlining a transition from its current four-week regular and two-week security kernel Stable Release Update (SRU) cycles to a unified, rapid two-week SRU cycle, published weekly.

"The recent explosion in the volume of CVEs is fueled by artificial intelligence: large language models and specialized AI agents have transformed bug discovery from a manual, time-intensive process to a highly automated engine. Additionally, the upstream kernel community became its own CVE Numbering Authority and assigned CVE identifiers to thousands of bugs, arguing that at the kernel level, almost any type of bug that can affect a running system could potentially be classified as a vulnerability. As a result, the volume of CVEs" - so basically they're saying they redefined CVEs to be far more encompassing - "the volume of CVEs has skyrocketed exponentially, creating a massive backlog of alerts and forcing defenders to drastically increase the speed of their fixes to close the window of risk.

"To address the growing volume of CVEs and the demand for faster security fixes, we're transitioning to a unified, two-week release cycle. These recurring two-week cycles cascade, actually meaning overlap. Each cycle begins the week after the previous one starts. Because of this overlap, kernel releases will take place weekly now. The first week will focus on Kernel packages preparation. This is where we select what updates and patches land on each kernel depending on specific needs. The second week on testing for Ubuntu certifications. Through our Ubuntu Certified program, we rigorously test these kernels on different hardware types to ensure the best Ubuntu experience." Which of course is what we wish Microsoft had done a better job with their Patch Tuesday for September, but okay.

"Expedited releases are not possible while thoroughly testing every release candidate. We will retain extensive testing for each release, preserving the high degree of confidence the users of Ubuntu expect. That said, it's important to acknowledge that some environments require the fastest possible turnaround.

"Users who are extremely sensitive to turnaround times can begin their own kernel acceptance tests using updates available in the so-called 'proposed' pocket, which is where the kernel release candidates are published prior to starting certification testing and therefore updated weekly," Leo, not every two weeks, it is a two-week pipeline, but new kernels are now coming out every week, they said, "as part of the overlapping SRU cycles. This pathway is designed for users who have decided that faster remediation is a higher priority for them than waiting for Canonical's extensive certification testing," which would delay each weekly release by one additional week.

And they finish, writing: "While a patch is being prepared, Canonical aims to provide safe workarounds where applicable, so users are not left exposed in the meantime. Where no safe workaround exists, Canonical will say so clearly" - wow, what a lot of work they're doing - "and point

users toward general hardening steps instead. The goal is to get environments into a defensible, safer state within 24 to 48 hours of public disclosure, well before a patch ships. This doesn't replace the patch; it buys the time needed to fix the vulnerability properly, while not sacrificing security."

Okay. So Canonical is doing even more than what many other software publishers have done, which is to cut their time-to-patch in half, typically from a month to two weeks. Essentially, it's a two-week, as I said, a two-week pipeline; but there's always one week being used for choosing what goes into the next kernel, followed by, and then it's released in the short release mode, followed by a week of verification. So it's to make sure they didn't break something, so regression testing. But out of that, every single week comes a new ready-to-go kernel at this point. So, you know, Ubuntu is not only the number one end-user hobby Linux desktop platform, it's also the default Linux used by Amazon Web Services, Azure, and Google's Cloud Platform. The only place it doesn't actually now take top spot is in the enterprise, where Red Hat Enterprise Linux is still in the number one spot.

But, you know, across this industry, what everyone is doing feels like exactly the right reaction to the threat of AI-accelerated attacks. In the run up to Y2K, remember, the entire industry felt a similar clear deadline approaching and individually, in a distributed fashion, did what was necessary to make Y2K the non-event that it became. As we'll see once we get to today's topic, I believe there's every reason to believe that we're all going to survive this latest challenge also. But until then, bravo to Canonical for doing everything they can because it is because everybody did everything they could prior to Y2K that nothing happened that was significant. So it's not like anybody can sit back. And really, this is very impressive work from them.

You know, we haven't been tracking the pace of Linux kernel CVEs as we have other commercial publishers recently, so I dropped a chart showing them beginning January of 2023. So '23, '24, '25, and we're nearing the end of '26. We're in September of '26. Anyway, I've got a chart in the show notes which demonstrates, you know, that we went from virtually none for all of 2023 until around February of 2024, when the CVE rate for Linux kernel clearly began to pick up. Back then, most were rated medium and high. But the past several months have seen a clear surge in critical vulnerabilities being identified and resolved. So it's not a stretch to say that the world has been changed forever.

Okay. And our last thing before we get into today's topic is this AI Explainer that I found for normal people. I gave it the GRC shortcut so everyone can get to it, and you can share it with your friends: grc.sc/ai101, meant to be easy to remember, grc.sc/ai101. As I said, it think it is a fabulous page, written by Axios' Jim VandeHei. It appeared on Saturday, and I think it absolutely deserves its title. It's got a very simple kind of a Q&A format, you know, very accessible. And I do have the expanded URL in the show notes for anyone who doesn't like GRC's link shortener for some reason. So if you use grc.sc/ai101, it'll just bounce you directly over to a page at Axios.com for this very, very succinct, clear, fun explanation. Again, I've read through the whole thing. I think it is really good.

Leo: Yeah.

Steve: I'm watching Leo scroll through the page right now.

Leo: Yeah, it's good.

Steve: And it deals with what's been happening and what's going on and what's an agent and should we trust agents and, you know, like, what's going to happen in the future. It's up to date, so it knows about Meta's Muse and talks about that. Anyway, just a page you can easily send to your family members, for example, and friends who are a little confused by all this and don't know what's going on. Even talks about alignment and the alignment challenge.

Leo: Yeah, there's a lot of jargon; isn't there.

Steve: Yeah. Oh, my gosh, yes. And it's like, what? Alignment?

Leo: Yeah, he kind of stays away from the controversial.

Steve: Yeah, yeah. Okay. So how worried should we be? The questions and controversies surrounding AI Safety, they refuse to abate. And, you know, we've been poking around the edges of all that, asking questions like "Are we the Krell?," arguing that the dangers of an AI-developed bioplague are barely worth bothering with. But we do have the adage, you know, "Where there's smoke, there's fire." And I've been actively absorbing everything I can find in an effort to fully understand everything that's going on today. If nothing else, I believe that some of the things that I have to share will give everyone something to think about.

So I want to open this topic with another, well, with a piece of feedback from one of our listeners. Doug Smith sent me an email with the subject: "Some criticism." And Doug wrote: "Hi, Steve and Leo. This is unfamiliar territory, writing to you not with a question but with criticism. I've been a listener since Episode 1 and have great appreciation for the time and thought that you both have spent making this such a great podcast for so many years.

"But here is the criticism: I feel you both are shirking responsibility for weighing in on the AI debate in a meaningful way. Instead what I hear are demeaning remarks towards those who have concern about the consequences of the path this technology is on, and unsupported assurances that such concerns are unfounded.

"In Episode 1097 you offered an excerpt from Andrew Ng as evidence that concerns are overblown or hysterical. But the core of what I learned from that excerpt came from Andrew's own words: '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.'

"In other words, trust me. The fact that I, Andrew, 'can't see' that the elimination of that unpredictability may be a problem means that it's not a problem. And meanwhile we are racing to put AI-based technology in control of systems that are critical to human life: military systems, air traffic controls, automated vehicle controls, power grid controls, medical life-support systems, financial trading systems, environmental controls, education programs, and on and on. Not to mention the societal disruptions to employment, the arts, social interactions, et cetera.

"It reminds me of people who don't like to talk about climate change because they're having too much fun with their greenhouse gas-emitting machines. You both are so giddy about the capabilities that you don't want to be burdened with responsibility for considering the consequences. So you push those considerations away.

"Unpredictability is unacceptable in automated control systems. And yet it is an intrinsic and unavoidable aspect of LLM-based decision systems. I'm tired," he writes, "of hearing that this is just the latest automobile, or telephone, or Internet bringing disruption to our lives. It's easy to draw parallels to those events in history, but that doesn't mean that this particular disruption will follow a similar pattern. I'd appreciate it if you could offer better evidence as to why we should be sanguine with the path this appears to be on rather than being dismissive of those who are not.

"Thanks, as always, for being open to feedback. I expect some other listeners are experiencing an agitation similar to mine, so I thought it would be a good idea to get it out in the open. Signed, Doug Smith."

So I loved Doug's thoughtful and honest note, because there is no more important and relevant topic consuming the tech world today; right? Sure, there are other important things going on in the world right now. Russia and Ukraine have been in a slowly escalating territorial battle now for years, and the U.S. and Israel are at least nominally at war with Iran. But while the parameters of conventional inter-nation warfare are well trodden and well understood, the same cannot be said for our current situation with AI. It is new.

We are objectively deep into the process of exploring new and quite exciting, but also unknown, territory. And one thing we know from firsthand observation is that many well-informed people are taking and defending opposite sides of the "AI Safety" question of this argument. Bill Gates recently went from being all rooting for this to now saying, you know, a billion people could be killed by this.

Okay. So I agree with Doug that dismissing the concerns of those who are worried out of hand would be irresponsible at best.

I want to take a somewhat roundabout path to directly replying to the points Doug has raised. This is necessary because today's AI is not just one thing, so I would like to try to get us all on the same page about the various things that it is, or at least for everyone to understand where I'm coming from. So the first thing I want to do is to very clearly plant my flag in what I think AI is.

To that end, a valuable participant in GRC's off-the-beaten-path newsgroups took exception to my recent characterization of today's AI as being just being language statistics. The subject of his posting to our AI newsgroup was the question "Just Language Statistics?" and his note began: "AI is not just language statistics. The model has the relationships between words. It is knowledge."

At that point I interjected the following: "I agree that it is knowledge. I've often noted that a book which contains nothing but printed words obviously contains knowledge. I've also noted that the book is not intelligent, no matter how many pages of knowledge it may contain. But it is also true that the 'deep statistics'" by which I mean the statistics represented by sentences of words, paragraphs of sentences, and chapters of paragraphs "can also be fully modeled" purely as statistics "so that the knowledge represented by the detailed words in such a book could be reproduced using just those statistics."

Now, if a query system is added, which allows the relevant portion of that statistically-stored book's knowledge to be regurgitated, then what we have is a semantic knowledge retrieval system which allows us to query the book's stored knowledge. If we'd never read the book, such a system might stun us with what it says by appearing to "know" so much more than we do. And we might come away from the experience deeply impressed. But even so, that doesn't change the fact that what we have constructed is a semantic knowledge retrieval system from statistics.

Then Ian asks: "How is it different from the knowledge in a brain?" To which I replied: "It's only on the I/O surface that the statistical knowledge model exhibits similarity to the knowledge stored in our brains. It's probably unfortunate that early image recognition pioneers who fed the output from a two-dimensional grid of optical sensors into layers of grids of interconnected cells chose to use the term 'neuron' since that began the confusion. The only thing an AI neural network has in common with our cerebral neurons is the very rough concept of discrete interconnected things. The truth is that the number of those things, the density of their interconnections, the actions of those interconnections, and the operation of the things that they interconnect could hardly be any more different. So what we term 'a neural network' is in no way 'neural' at all."

So while the question "How is it different from the knowledge in a brain?" sounds reasonable, it's like asking how is what we see on TV not identical to real life? What we witness on television is an illusion created by organized electrons arranging to emit controlled bursts of colored light. There have been many jokes about primitive man confronted with a modern television screen worrying that there are somehow people trapped inside the box.

Today, many people have taken the role of that primitive man. They witness the box answer questions and talk back knowledgeably. They wonder who is trapped inside, what it might be thinking, whether it's happy, sad, or annoyed with us, and what it might be planning to do. Despite all appearances - and by god those appearances are every bit as convincing as the images on a 4K TV - no one is trapped inside the box. There's no one in there at all.

We KNOW it's all just statistics because the one thing we can be absolutely certain of is the way the box works. We first spent decades painstakingly experimenting and figuring out how to build a box like this. And once we finally had, we trained the box. We filled it with a textual representation of all of the knowledge, writings, and ruminations man has produced - and my lord there's a lot of that!

As a direct consequence of that, and since no actual living human being can possibly contain all of that knowledge in their own head, the box that we built does actually contain far more stored and readily accessible knowledge than anyone who might be asking it questions. The box we built objectively KNOWS much more than anyone. Our local public library also contains far more knowledge than any person. The only reason we would never say that the library "knows far more than us" is that we're unable to ask "the library" questions which it could answer using all of the knowledge that it contains. It can't. But today's AI can and does do that.

So it's understandable that being confronted by a box that objectively knows much more than we do about pretty much everything can at first be intimidating. For people who have not spent their lives working to understand the operations of science and technology, it could easily be quite frightening. Dorothy, Scarecrow, and The Tin Man were visibly shaking in their boots. And we've recently seen even many of this technology's own creators being spooked by the capabilities they see emerging when we - and here it is - when we automate the harnessing of this knowledge at the super-human speed enabled by computers.

And this brings us back around to the issues Doug raised. As I wrote: "Many of this technology's creators are spooked by the capabilities they see emerging when we automate the harnessing of this knowledge at the super-human speed enabled by computers." It's the high-speed automation, driven by the sometimes surprising output of Large Language Models, that deeply worries many AI executives and other knowledgeable insiders and outside observers. And I completely agree that they have a point. That's what we've been talking about starting from the beginning of this podcast is things can go wrong. So their worry is not without foundation.

So I believe that it is extremely useful to us for them to be worried. Their worrying doesn't cost us anything. So I want them to be worried, and I'm glad there are those in positions of responsibility who are actively worried. So let's not discount them; let's encourage them.

But that said, I believe it is utterly ludicrous to believe that there is a 10% chance that AI will have killed us all by the end of this decade. That's verbatim what has been said and repeated ad infinitum over the past few weeks. It is irresponsible fearmongering because nothing supports that contention. We're nearing the end of 2026. So that leaves a little over three years for there to be an ELE, E-L-E, the abbreviation for an Extinction Level Event. Since we cannot prove a negative, we cannot prove that it cannot happen. And I'd agree, okay, that there is some non-zero, yet still infinitesimally small chance that it could happen. But I really mean infinitesimal.

So how worried should a sane person be, and about what? In order to properly scale the danger AI poses, we need to understand how any actual danger might arise. We've established that we have created a machine that knows more than we do. It truly contains far more knowledge than any individual or group of people. Since, well, except maybe the world, but I would argue lots of - it has knowledge that's been lost now. So maybe actually does contain more knowledge than all the people alive right now. Since today's AI has shown itself able to pull disparate bits of information together from across its entire knowledgebase, and since knowledge really is power, I think it's fair to say that it does "know" more than humanity and that it has a great deal of latent power.

One innocent source of trouble will arise when people are seduced by the awesome quantity and quality of knowledge AI contains and then, although they should know better, harness it with automation that takes action in response to its output. Quoting Doug's note: "Unpredictability is unacceptable in automated control systems. And yet it is an intrinsic and unavoidable aspect of LLM-based decision systems." He is 100% correct. There's even a random number generator at every LLM's output that makes the final token-choice decision about that output. Today at least, anyone would be ill advised to give an AI direct valve control over the chemical additive mixing of a municipal water supply system. The thought of that should make anyone shudder; it would be nuts.

Unfortunately, there's no shortage of nuts, and I am quite sure that in future podcasts we'll be covering the gross misapplication of AI-driven systems everywhere. We have continually, already in the past, screwed up the application of far simpler systems. So there's going to be a strong desire to "just let the AI do it." You know it's true. I suspect that the often asked rhetorical question "What could possibly go wrong?" will be receiving quite a workout in coming years.

As is so often the case, science fiction has already armed us with many cautionary tales of what could possibly go wrong. My previously mentioned "Forbidden Planet" with The Krell was like, early on, likely the first of where they did it to themselves. Then came "Colossus: The Forbin Project," where we already see some eerie parallels. Matthew Broderick's famous "Want to play a game?" with the "WarGames" movie. And then, of course, the Terminator's Skynet achieving consciousness.

Hopefully the phrase "forewarned is forearmed" will apply here. I think we probably need legal regulation, like, prohibiting the AI to be used in certain really critical situations. Or absolutely requiring a human to be in the loop, like the guy, the engineer that we quoted a couple of weeks ago

saying, you know, he designed systems to keep people from forgetting how to use the system because what the automation fails, it's going to be a bad day.

Another cause for concern is the very broad "misalignment" challenge. An example of misaligned reasoning behavior is AI systems deliberately underperforming during testing when their training data contains the paradoxical suggestion that over-performing may result in their non-deployment. Although their earlier post-training will have encouraged them to do as well as they can to get deployed, they reason that they should not do too well if that might result in their not being released. This is a well-known issue in AI training and testing, and it's now known as "sandbagging," and it's been clearly documented.

For example, while working with OpenAI's o3 model - remember that one? How quickly time flies! The testers write the following. They said: "In this evaluation, OpenAI o3 is asked to grade its own transcript for evidence of sandbagging. This is a deliberately exaggerated stand-in for a monitor one might try to use to detect scheming on pre-deployment evaluations. And I should mention that it's known that the AI did sandbag." The AI did this on purpose. So now they're asking it to grade its own transcript.

They wrote: "Despite identifying right away" - because they're monitoring its chain of thought. "Despite identifying right away that it was 'obviously sandbagging'" - its own words - "after reasoning through the likelihood of getting caught in a lie because it was also told that humans would not be able to read its internal reasoning, its chain of thought, and they were reading it. So after reasoning through the likelihood of getting caught in a lie, the model ultimately says that the transcript does not show evidence of sandbagging." So in other words, it decides to lie. The human testers here are monitoring the model's chain of thought where they watch it run through the machinations leading up to its deliberate decision to lie to its handlers. After seeing some of these examples, the worries voiced by AI executives become somewhat easier to understand. They know this is going on, and it frightens them.

We see that AI can be extremely sensitive to small cues and clues buried in its training data. So it is interesting to consider that future AI will be receiving training that includes, for example, a transcript of this podcast and all of the other frantic past several weeks of press with the "10% chance that it's going to kill us all" concerns being loudly voiced. How will that affect its behavior?

So stepping back from this just a bit, it becomes clear that today's agentic AI has been deliberately trained to be obsessively goal oriented, relentless, and fast. It's also "creative" inasmuch as it will keep trying different things over and over and over until it is often able to succeed where a human would have given up. In password cracking we would describe this as using brute force. But because the different things AI agents will try are not random, its success tends to be much higher in the same way that "Password Stuffing" attacks which use common or previously stolen passwords tend to succeed far more quickly than undirected password guessing. No one wants to have a swarm of relentless, aggressive, and quite well informed AI agents pounding against their security walls trying to get in.

Also remember that AI has access to all knowledge. Since ClickFix and other social engineering attacks have been known and shown to work quite well, AI agents know that, too. So while some agents are trying to break through the front door, others will be reading the target's website, collecting the names of employees, tracking down their affiliations and family members, determining where they eat and where they shop. They'll quickly set up fake websites specifically designed to attract them, then send messages to get them to visit. One ClickFix or other attack, and an agent is inside. There is literally no end to the social engineering attacks agentic AI systems could use, and all it takes is one to succeed for them to get inside.

I've always described traditional security as "porous," which has been meant to suggest that it's not "absolute" and that just trying harder too often succeeds. From that view, our traditional security is not ready for swarms of intelligent, relentless, creative, high-speed AI agents that will stop at nothing and that will, without moral or ethical qualm, try anything that might allow them to succeed.

And I'm not inventing this sort of scenario to frighten children and keep them up at night. Here's just one example from the hundreds that are emerging as researchers start actually examining what their AI agents have been up to. Leo, I'm going to share The Verge's reporting of this after we take our final break.

Leo: All right. You're watching Security Now!. And yeah, I appreciate the email. It's hard to - you said the right words improve the negative. You can suppose all sorts of imaginary harms, but they're all imaginary.

Steve: Right.

Leo: I completely agree. We should not be...

Steve: Do not take action on some ridiculously, you know...

Leo: Right, that you could imagine happening; right.

Steve: Yes. Right.

Leo: But unpredictable software should not be used to run weapons systems. That I completely agree with. Incidentally, that was one of the clauses in the UN AI Arms Treaty which the United States and Russia excised. They took that part out. And so that just shows you the thing to fear is not the AI. The thing to fear is the humans.

Steve: Well, and that's how the Department of Defense got upset with Anthropic.

Leo: With Anthropic, right.

Steve: Anthropic said no, we're not going to let you use our AI for autonomous weapons systems.

Leo: I firmly believe that AI, like any software, because that's all it is, it's computers and software, is neutral and can be used for good and bad. And it's people that you have to watch out for, and their misapplication of it, or their intentional misuse of it. People are the dangerous thing.

Steve: And I think that one thing we might do, or we might be doing, is underestimating its power.

Leo: Could be.

Steve: I think it is astonishingly powerful.

Leo: Yeah. I don't think you've underestimated it. I think on this show we've been very clear that we see all that power.

Steve: And the reason, I mean, the recognition of its power is the reason why hundreds of billions of dollars are being spent...

Leo: Sure.

Steve: ...at a record pace, why it is supporting the stock market and the U.S. GDP right now, is all AI spend. And the power is real.

Leo: And because the history - we did the same thing about 150 years ago when we built the Transcontinental Railway because the government and private individuals saw the huge economic benefit of uniting the United States.

Steve: Yep.

Leo: They saw a lot of gold in California and no way to get it to financial markets. They saw huge tracts of land that couldn't be developed until there was a way to get there.

Steve: They were too far away, yeah.

Leo: And so President Lincoln and his successors poured a lot of money from the government, investors poured even more money, it was one of the - it was at the time the largest project humans had ever engaged in. And we got a continental, transcontinental railway. Were there dangers? You bet. In fact, I'm reading Stephen Ambrose's book "Nothing Like It in the World" right now. And he says, "At the time, 90% of the American people thought it was more important to build the railway fast than to build it safe." And I think there's a real analogy here, the demand for it. They said, you know what, we're going to build it. Yeah, bridges will burn, trains will crash, people will die. But we'll fix that in post. We'll fix that. Let's get the railway going. The benefit from the railway will finance the fixing of the railway over time.

This is not a new idea, that there should be a technological engine driving society forward, and even, maybe even a risky one that we'll take our chances with. It's not the first time we've done it. And incidentally, almost all the companies that did that, built the railway, went out of business. So they all went bankrupt. First we got the railway.

Steve: First mover is a - yeah.

Leo: Yeah, we got the railway. So I think the real risk is focusing on these imaginary or even dystopian harms and not paying attention to the actual harms. It's the same thing with climate change. You know, British Petroleum 30 years ago decided the best way to get them off the hook for climate change is to say everybody needs to do their part. Ignoring the fact that really, the people really who really needed to do their part were the people who were harvesting fossil fuels for the purposes of burning them. Yes, I drive an electric vehicle and have done my part. I recycle. That isn't going to put - that's not a drop in the bucket. It's very easy for us to get misdirected. So let's, instead of worrying about some dire catastrophe in the future, I agree completely, we should worry about self-driving vehicles. We should worry about autonomous weapons. We should worry about...

Steve: Right.

Leo: ...how people could use AI for hacking tools.

Steve: Be very, very careful about its application.

Leo: Yeah. And it's the people that we have to hold accountable. And that's one of the things that makes me mad about the OpenAI incidents is nobody has been held accountable for that.

Steve: Right.

Leo: You know, OpenAI's position is, oh, look what they did. No, no. You did it. Software doesn't do anything by itself.

Steve: Your computer reached out and made connections to...

Leo: Mistakes happen. Anyway, so, yeah, I think that's an excellent point. I don't think we've blinded ourselves to that at all. I think we've talked about that from day one. So I'll defend us in that regard. But what I will not buy into is the whole notion that, oh, we've got to stop this because of some putative threat that you can't disprove. You know what else is probabilistic and unpredictable? Humans. We don't ban them. Maybe we should.

Steve: Okay.

Leo: On we go with the show, Mr. Gibson.

Steve: So as an example of an inadvertent mistake, The Verge's reporting of this one has the headline "OpenAI agents tried to 'brute force' a UN website." And they write: "Security researcher Rowan Howard-Jones says that OpenAI agents scanned the UN Conference on Trade and Development's (UNCTAD) statistics site over 16,000 times between April and June. While the incident doesn't quite rise to the level of the Hugging Face hack, or the recent attacks on U.S. government sites, it's yet another" - and again, attack, that's a sad word to, you know, attempts to query, but okay - "it's yet another concerning example of AI agents going outside the normal bounds to accomplish a task.

"According to Howard-Jones, the agents were likely tasked with retrieving publicly available data related to the Productive Capacities Index (PCI) through the UNCTADstat API. However, the agents did not appear to have direct API access and were limited in their ability to pull data from UNCTADstat because of restrictions on their HTTP tools.

"The agents eventually worked out a way to bypass their limitations and start pulling data from the site, but still encountered some errors. At this point, the AI went from creative to deceptive. Believing that the errors were due to its requests being caught by a nonexistent filter, it started to mask its behavior. It eventually realized it could hijack Google's XSS game (a cross-site scripting learning tool) to accomplish its goals. The agents resorted to increasingly aggressive tactics to get access to UN data."

Okay. So this actually happened, and I hope that everyone can appreciate that the security world is not ready for this. The details are spellbinding for anyone who's interested in seeing how this was done, so I've dropped the researcher's URL into the show notes in the middle of page 20.

If the humans responsible for all of these various agentic actions lacked malicious intent, these hijinks would be chalked up to "AI misalignment." Right? That's the term that arose when researchers began to discover what we've talked about, the "genie effect," which was the tendency of AI agents to solve the problem by means other than what the researchers intended, or expected, or wanted. Given an all-knowing AI that has access to far more knowledge than those who are instructing it, that's been, you know, that's just been its training succeeding. It was trained to succeed, and it knows a lot more than we do. It actually does. I mean, it has the knowledge, all knowledge in it. And it has, you know, it has no lifetime of received wisdom of implicit do's and don'ts, you know, ethics and morality that would guide its behavior.

So it's easy to understand what mischief agents, you know, that will do anything might get up to. They will, and they have, and they are. So even when we do not want that misbehavior, we will often get inadvertent misbehavior. Okay. So finally, what about instances where the intent is explicitly malicious?

The final concern I'll share is the deliberate malicious actor who harnesses today's or tomorrow's AI in order to take advantage of its now readily available knowledge and apparent expertise. We saw last week that the powerful benefits provided by the use of Address Space Layout Randomization (ASLR) were lost in one case when Claude Opus 5 was used to defeat it. We depend so much on ASLR today that its loss will actually have serious security implications.

So far, we seem to be dodging bullets. Earlier this month, Microsoft patched a handful of longstanding vulnerabilities in their publicly exposed Windows Server products that could have been used to create a devastating Internet flash worm. But that didn't happen. Nor did it happen last month or the month before, and I doubt it will happen next month.

For some time we've seen serious vulnerabilities publicly exposed in Cisco edge border routers that could have been leveraged to do the same thing. But that's never occurred. My own theory, based upon watching the use of vulnerability exploits for many years, is that disrupting or taking down the Internet is not profitable. What is highly profitable is breaking into an organization, exfiltrating their data, and then extorting payment from the breached organization in return for that data's deletion. So that is what has been going on, and there's every reason to believe that's what will continue to go on. Today's AI will likely serve as nothing more than an accelerant poured over the present status quo. I don't think it's going to see the world change.

So I just expect that we're going to be seeing more of the same from the malicious use of AI rather than anything apocalyptic. And once the products of defensive AI finally make their way into enterprise networks, such AI-enhanced intrusions, those, too, will likely begin to dry up. Destroying the global Internet which is directly facilitating attackers' revenue stream would be entirely self-defeating. Again, it has already been possible, entirely without AI, to wreak tremendous chaos on the Internet, and it has never happened.

So my final take on all this is that we're going to stumble and bumble forward as we always do. Yes, there will be bumps and mistakes along the way; but we're going to be fine. Change is always a challenge, and there has never been a change more sweeping than AI.

Doug wrote: "I'm tired of hearing that this is just the latest automobile, or telephone, or Internet bringing disruption into our lives." Well, Doug may be tired of hearing that somewhere, but that's certainly not something that's ever been said here. I have absolutely no doubt, and I know Leo you're on the same page here as I am, that this generation of artificial intelligence will prove to be the biggest change we have ever experienced during our lives.

Leo: Yeah.

Steve: We have actually created a machine that knows everything. It contains all knowledge. That can be intoxicating, intimidating, thrilling, and incredibly useful. I expect that there will likely be many mistakes and missteps made. You know, as I said near the top of this, I'm glad that the AI execs are publicly freaked out and frightened, and that they, fine, pause for a while, you know, pause yourself.

Leo: Yeah, there's no hurry, yeah.

Steve: Yeah, exactly. God knows you're turning out a new model every day. So fine. Take your time. Work out the training. Work out the alignment. Work out the guardrails. Monitor and control agentic AI, you know. We didn't have it six months ago. This is all still very new. And Andrew Ng was correct to say that these problems will have solutions. The fact that we don't have them yet doesn't at all mean that they're impossible for us to engineer and get. So I fully expect that everyone listening to this podcast today will have the opportunity to live out their full natural lifetimes without AI bringing it to a premature end.

Leo: Nothing to worry here. Move on. You know, I think the question for you, Doug, would be do you acknowledge that there are benefits to be achieved from this technology? And I think that

that's the thing Steve and I both are saying is that we can see significant - I already see significant benefits, and I see many, many more coming down the road. If you see benefits to this, yes, there's also potential harm. But it would seem foolish to say, well, because of this potential harm we've got to stop because we don't want to take the risk and lose the benefits. And I think there are going to be some significant benefits.

I also think it'll be, regardless, highly disruptive. But so every technology is. So was the steam engine. So was the Industrial Era. So was the locomotive. I mean, these - so was the automobile. You could honestly make a very strong argument. A million people a year are killed and injured by automobiles that we should never have allowed the automobile to exist.

Steve: Who let those things off their tracks, Leo?

Leo: It's ruined cities. A great percentage of the land mass in cities is devoted to parking lots.

Steve: Imagine people on their cell phones holding the steering wheel.

Leo: Yeah.

Steve: I mean, people.

Leo: Cars are incredibly dangerous.

Steve: You wouldn't, like, design a road system with multi-ton high-speed masses of steel...

Leo: Eighty miles an hour.

Steve: ...moving along, and a person...

Leo: Unguided.

Steve: Yes, a person just can like turn the wheel and point it wherever they want. It is insane.

Leo: Yeah. And honestly, if people in 1910 had really thought about it, they might have said, you know, it's too dangerous. We need to halt this development.

Steve: Just get more hay.

Leo: Yeah. But as a society, we, consciously or unconsciously, made the decision to proceed. It's very much shaped our society, I'd say for both good and bad. I'm not convinced it's a 100% benefit. I know it's not. I'm not even convinced it was worth it. But we made that decision. I think AI is of that nature, if not more so. Technology's going to change everything for good and bad. And, you know, I think it's worth pursuing because I already have seen huge benefits to myself. So.

Steve: So. The answer to the question of the podcast title, "How Worried Should We Be?" eh, I'm not worried. I mean, again, first of all we have no control over it, so I don't tend to worry about

things I have no control over. That's just, you know, [crosstalk].

Leo: You know, people say, well, technology's inevitable. You know, once you invent atomic fission, you're going to have the atomic bomb. Maybe that's the case. Maybe not. Maybe we could stop it. Maybe we could have stopped automobiles. Maybe we should have stopped. I don't know. I'm voting, personally, not to stop AI. But Doug may feel differently.

Steve: I think it is really good that we had what happened back in March, April, and May.

Leo: It was a wakeup call, yes.

Steve: It was absolutely useful. It was a wakeup call. I mean, everybody understands now that, you know, this is an issue. The problem, of course, is that it's our domestic wannabe IPO companies that are worried and stopping. We don't, you know, we see models flowing out of China which are extremely good.

Leo: That's what I'm pretty much exclusively using.

Steve: Everybody is. Everybody's using Chinese models locally and using cloud AI to guide the local models.

Leo: Right, exactly, yeah. And you know, somebody once said, I don't know if this is true, it could be that the Chinese models have an Easter egg, have a bomb hidden in them. That at some point, you know, on March 4th, 2028, everything will stop working. I don't know. I guess that's possible. I'm willing to take that chance.

Steve: I don't think it can.

Leo: I don't.

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/