AI-Driven Expertise Loss

Description: We start out with a classic old-school hack against Dropbox. Next Patch Tuesday will be enabling "Memory Integrity" for many. Firefox moved to 155 and obtained a dumb Smart Window. CISA is terminating six most valuable cybersecurity services. OpenAI advanced to top of the heap with GPT-6 Astra. But is it now hiding some of its thinking from monitoring? NVIDIA is acquiring Hugging Face. Who's that good for? Google releases Gemini 3.8 Flash and Cyber. Is it good? Chinese cyberespionage is using AI to become more slippery. Matthew Green proposes a fascinating take on AI bug drought. A lifelong safety engineer contemplates AI-driven loss of expertise.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1095.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1095-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. It's Patch Tuesday, and you won't believe how many patches Microsoft just shipped. Steve says that's okay. They're fixing things. New versions of Google's Gemini, NVIDIA acquiring Hugging Face, Chinese cyberespionage, and then a lifelong safety engineer worries that AI is going to cause us a loss of expertise. That and a whole lot more, next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1095, recorded Tuesday, September 8th, 2026: "AI-Driven Expertise Loss."

It's time for Security Now!. Yes, Tuesday has come around once again, and that means Steve Gibson is knocking at the door, ready with a 22-page document of all the latest security problems in the world. Hello, Steve.

Steve Gibson: Yo, Leo.

Leo: It's good to see you.

Steve: For those who've looked at the show notes, Benito was the first person to highlight the fact that I had numbered this 1096. This is not 1096. This is 1095.

Leo: Fixing it right now, yes.

Steve: So the links are right, but in the show notes I had it wrong. So anyway, we are - this is - and I have not yet looked. We are on Patch Tuesday, and a big question that we have is how does this one compare to the last one? Which of course was a whopper.

Leo: Your theory is that it'll get better at some point.

Steve: I believe - I don't know what the shape of the curve is. I don't know how soon it's going to get better. But in theory, as long as AI is aiding us in eliminating more bugs than we or it is creating, we should be seeing a drop-off because we certainly are seeing that we're paying for a lot of the legacy code mistakes that have been made in the, you know, previous era.

So we're going to start out talking about a classic old-school hack against Dropbox that actually something happened that was not about AI, believe it or not. We've got, as I said, oh, the Patch Tuesday after today is going to be enabling something that Microsoft first - it first appeared in Windows 10. And we talked about it back then. The short name for it is Memory Integrity. Many people turned it off because it dropped gamers' frame rates in the interest of improving their security. A lot of gamers said, "I'm secure enough, I need a high frame rate." Anyway it's going to be turned on next month for everybody that qualifies based on the hardware. So we'll talk about that.

Firefox moved to 155 and obtained what I would call a dumb Smart Window. So we'll touch on that. CISA has terminated six of its most valuable cybersecurity services. And really you couldn't choose a worse time for CISA to back away from, you know, infrastructure cybersecurity because everyone's expecting AI to impact that. Well, everyone except me. I'm not convinced that that's going to happen. We'll talk about that.

I want to touch on OpenAI's big announcement because last week, since we talked last, lots of things happened. OpenAI has released GPT-6 Astra, which is that model, we did discuss it before, that they're saying qualifies for what they consider critical security treatment. And then also there's been a weird bunch of what I consider misreporting angst generated by the, you know, the anti-AI guys saying that OpenAI is doing something which is hiding its own internal thinking, masking its chain of thought. I'm going to open that up, take a look at it, and explain why that is not what is going on.

Also we'll touch on NVIDIA's acquisition of Hugging Face and who that's good for. Google released Gemini 3.8 Flash. You know, having waited, what, two weeks from 3.7. So this is all happening very quickly. Chinese cyberespionage is beginning to use AI to become more slippery. We'll talk about that. And Matthew Green, our favorite Johns Hopkins cryptographer, has a fascinating take on what it means for there to be an AI-driven bug drought, which - he has some very worthwhile thoughts. And then today's topic is "AI-Driven Expertise Loss." A lifelong engineer, who has designed the control systems for nuclear reactors, I think has some very important things to say about the consequence of AI being so good, and what it means in the long term.

So lots of fun things to talk about, and another great Picture of the Week. So, yeah, this is Episode 1095, despite what the show notes say. And we'll get to the [crosstalk] next week.

Leo: And do you want to know how many patches there are in the Patch Tuesday this week?

Steve: I do.

Leo: Are you curious?

Steve: Yeah.

Leo: Well, Tom Warren on The Verge says I understand today, September's Patch Tuesday, is about to set a new record with more than 650 security fixes for Windows.

Steve: Whoa. Whoa.

Leo: 650. Last month was 400. We thought that was a lot.

Steve: It was.

Leo: July was 570. We thought that was a lot.

Steve: It was a lot. Yikes.

Leo: 650. So it's not going down.

Steve: Not yet, no.

Leo: And Microsoft says, yeah, we're using AI. We're finding them.

Steve: Wow.

Leo: So there you go.

Steve: Well, I mean, and so it's a mixed blessing; right? It means that we've all been riding on this just Swiss cheese, how does it even boot operating system. Wow. That's great. Well...

Leo: So it hasn't - I don't think Microsoft has actually released it yet. But that's what Tom Warren, who's very well connected and usually very reliable, says that that's the number.

Steve: So a bunch of operating systems are going out of extended service life. Oh, no, no. It got extended.

Leo: That's right.

Steve: We have until October of '27.

Leo: Another year, yeah.

Steve: And that's really good because by then they should have Windows 10 finally bug-free because, you know, we are getting security patches backported to Windows 10. And at the same time, they're no longer writing any new stuff to screw it up. So it's perfect. This is exactly what you want. You want another year of fixes without them adding anything new that's going to destabilize it. And then when it finally goes out of service, if they actually do take it out of service next October of '27, it'll be a solid operating system.

Leo: So I'm going to correct myself. About 10 minutes ago the SANS Institute put out their bulletin, and they said September Patch Tuesday has 973 patches, including 113 critical patches. So we are in uncharted territory. Two vulnerabilities listed as exploited in the wild. None were publicly disclosed before today. Windows privilege escalation and critical RCEs in Skype for Business MSMQ and [crosstalk].

Steve: Nearly a thousand in one month.

Leo: Unbelievable. That is stunning.

Steve: Yeah.

Leo: God. Wow. But as we should emphasize, it's good they're getting fixed.

Steve: Yes, it is. Remember those days...

Leo: We want to find these.

Steve: ...we used to have 23, or 12.

Leo: And I thought that was a lot.

Steve: Those quaint days.

Leo: SANS Institute is reliable; right? I mean, if they say...

Steve: Yeah.

Leo: ...it's that, it's that. That's unbelievable.

Steve: Yeah. I will have full coverage of it next week. We will understand what the demographics of those 973 are, and the 113 critical ones, just how critical. So, wow. Anyway, I would say update Windows.

Leo: Update your Windows. Yes, sir. Yes, sir. Our show today, we'll get to our Picture of the Week. We've got a fun one for you in just a bit. And now back to Security Now!, and it's time for the Picture of the Week.

Steve: So this is an xkcd, one of our favorite guys. And I didn't give this a title because he had, and it was perfect. The title of this series of cartoons is "Why Asimov" - of course Isaac Asimov, the famous sci-fi author - "Put the Three Laws of Robotics in the Order He Did." This is just pure genius creativity in this particular one. So what we have in the left-hand column is Possible Orderings of the Three Laws.

So the classic ordering is, basically, and he sort of reduces it for size, "Don't Harm Humans, Obey Orders, and Protect Yourself." And of course the full reading is under no circumstances harm a human being as the first law. And so we have Obey Orders, and the fuller version is, you know, do what you are told so long as it doesn't conflict with the first law. And then the third, which he shortened to Protect Yourself, is Protect Yourself So Long As Doing So Doesn't Conflict with the First or the Second Laws. So sort of, you know, a clean hierarchy.

And so this classic ordering, as I said, is Don't Harm Humans, Obey Orders, Protect Yourself, with the earlier law preceding any of the ones that follow. And when in the proper order, as Asimov intended, we see that all of this exemplifies in Asimov's various sci-fi stories, his famous robot stories, and it results in what we call a "Balanced World," meaning everything works.

Now you reorder those. So, for example, switching the last two, we have Don't Harm Humans, then Protect Yourself, and then Obey Orders Is Last. And so in the little cartoon we see the human saying, "Explore Mars." And the little cart, the AI-driven, says "Haha, no. It's cold and I would die." Which we sum up as a Frustrating World as opposed to the Balanced World. Or we swap the first two rather than the last two, so that moves Obey Orders into the first place, Don't Harm Humans into the second place, and Protect Yourself in the third place. And we see little cartoons of robots running around and atom bomb explosions and missiles flying through the air, which he summarizes as the Killbot Hellscape. Because of course Obey Orders comes before Don't Harm Humans, meaning that if you told your AI to go do something bad, regardless of the consequences to people, it would.

Then we have the case of all of them being reordered. Essentially Obey Orders is first. Then Protect Yourself, which is normally third, is moved up to second place; and Don't Harm Humans is in last place. Well, of course that's not going to turn out well. Basically whenever Don't Harm Humans is not in the first place, you get yourself in trouble. So again, a repeat of the third instance, same cartoon, atom bomb explosions, missiles flying through the air, and we get another Killbot Hellscape.

The fifth one down the order has moved Protect Yourself to first place. Okay, that's not going to turn out well. Don't Harm Humans under that. And Obey Orders, again, is in last place. So here the cartoon shows the AI robot saying, "I'll make cars for you, but try to unplug me and I'll vaporize you." Because of course Obey Orders down at the bottom, Protect Yourself is given first place. And that's sort of what we've seen with some of the AI that appears to have escaped containment. And we call this one the "Terrifying Standoff."

And finally, the last possible reordering is completely backwards. The original laws were 123; these are 321. So Protect Yourself in the first place, Obey Orders in the second place, and Don't Harm Humans in the third place. And this is another one of the Killbot Hellscape results with atom bombs and missiles flying through the air and so forth. So anyway, just a fun take on Asimov's famous Three Laws of Robotics. Which were elegant and simple and worked really well, when you think about it. It's like, you know, you need to not harm people. And as long as you don't harm people, you should obey the orders that people give you. And you should also try to protect yourself as long as doing so doesn't conflict with the either of the first two.

Leo: I agree.

Steve: So clean, elegant, simple. Okay. So in the midst of all the AI cybersecurity-related news, which has been saturating this podcast because it should, I wanted to start out deliberately this week's podcast with a blast from the past. This bit of news brought a smile to me because, for a pleasant change, it has absolutely nothing to do with AI and, as such, it feels kind of warm, comfortable, and quite familiar. It's the sort of news that we spent the first 20 years of this podcast examining.

Okay, so what happened? Dropbox disclosed a series of security hacks occurring between August 4th and the 21st, and so last month, during which attackers gained access and downloaded the private, confidential, and one would wish secure data, but not so much anymore, belonging to nearly 5,000 Dropbox users. Now, as I said, refreshingly, there's no sign of anyone using AI anywhere; this was strictly old school. The hackers accessed user accounts by abusing Dropbox's integration with Lenovo's identification service. And when Lenovo was confronted with this news, they stated that a "legacy integration" between Lenovo ID and Dropbox "could be used to improperly authenticate certain Dropbox accounts." Uh-huh, right.

So those pesky old legacy integrations that always seem to be allowed to endure right up until someone figures out how to abuse them was the culprit here. It seems that Lenovo ID users not using two-factor authentication had no protection, which allowed attackers to register with Lenovo's ID service - now, get this - using the same email address as a targeted victim, then somehow arrange to bypass Lenovo's email verification process I can presume that's where the legacy part comes in then use the newly created account having somebody else's email address associated with it to then pivot back to the equivalent connected Dropbox account, thus appearing to Dropbox to be the legitimate Lenovo ID user.

So depending upon whose account was hacked and what it contained, you know, although this was not a sweeping scope end of Dropbox attack, still the consequences could certainly be quite devastating to the individual users who were affected. So yikes. Once again, legacy bad.

Okay. So I had mentioned about this Memory Integrity enablement happening, coming next month, next Patch Tuesday, for Windows 11. Microsoft last week announced that next month's October 13th Patch Tuesday would be enabling Memory Integrity, which is a security feature, as I mentioned, that first appeared back in Windows 10. What's significant is it will be enabled by default on all Windows 11 machines. And this will not happen for otherwise qualifying machines if the machine's user had previously deliberately and manually tweaked the registry, or if the registry tweak occurred through a machine like an enterprise policy.

So as I said, we talked about this Memory Integrity feature back when it was first introduced, and as part of what Microsoft called Device Guard. It is a very slick system that takes advantage of the multilayered address translation hardware which is present in modern CPUs because the multiple layers of address translation also have privilege bits associated with them. So what we have with this Device Guard is what's known as "second level address translation," which allows a hypervisor which needs to be turned on and running. So that's - there's a little bit of overhead to doing this, which again is one of the reasons that gamers who are all frame-rate crazed deliberately disabled this when it began to creep into their systems, and people said, hey, what happened to my game play? It's not as good as it used to be.

So there's a hypervisor involved which sets physical page access permissions, you know, like writing to the page or executing code from the page. And it's able to do that separately from and effectively underneath the operating system's own virtual memory paging tables. So what's cool about this is it leverages the hardware from Kaby Lake on, over on the Intel side. I'll get to the specific hardware issues in a second. But it does so in a fashion that really strongly prevents a bunch of traditional problems. So as I said, first of all, on earlier hardware, there was more overhead than there is on later hardware. So because Microsoft decided this was so cool they were going to, like, do some emulation, which is never good because on the older hardware, this memory integrity being enabled noticeably reduced the system's, among other things, maximum display rendering frame rate.

Now, and the reason for that, as I said back then, and this was a decade ago, so quite a while back, was that the memory management at that time, the memory management hardware only had a single bit for controlling access at this second paging level stage. And Microsoft needed two bits. But there was only one. They need two for each of the possible modes, user mode and kernel mode, and be able to enable and disable those individually. So at the time Microsoft had to dynamically switch management tables on the fly anytime there was a switch between user and kernel. Hardware interrupts made that happen. Driver calls made that happen, and so on. So the overhead was very real, and among those who were really pushing their machine to the limit are gamers who noticed the difference.

But as I said, all that changed later with Intel's introduction of what they called MBEC, mode-based execution control. And that first appeared in the Kaby Lake processor family. AMD called theirs GMET, and that arrived in their Zen 2 family. And that changed everything. It removed the need for Microsoft to be doing any context switching, essentially, switching management tables as any time you did a kernel transition back and forth between the kernel and the user. There's still a tiny bit of overhead because, as I said, there is now a hypervisor active with the second-level address translation. So larger paging tables will inherently mean more paging table cache misses. So, now, this will increase cache miss rate, and so there will be some impact.

But modern hardware almost completely hides the overhead. And it really does return meaningful security. So a month from now, October 13th, any machine that did not have memory integrity enabled will have that happen to it. So if something, if like things seem to go slower after October's Patch Tuesday, it's not your imagination. But, okay. So here's what's actually going to change. Memory integrity has always been enabled by default for clean Win11 installs. And so-called "secured core PCs" like, you know, that ship with Secure Boot enabled. And that's been true ever since the Device Guard days. And that's why, because it's shipped enabled, gamers were frequently turning that off and seeing a performance boost.

So what happens next Patch Tuesday is that this is being extended, that is, from only clean Win11 installs and secured core PCs, it's going to be installed, it's going to be extended rather, to any

machines that may have been upgraded to Windows 11 from presumably Windows 10, or maybe he jumped over that from 7 or 8, and any that may have fit with it disabled, with Device Guard disabled. So you need to have qualifying hardware for this to happen, an Intel 8th Gen processor or newer, an AMD Zen 2 processor or newer, or if you have a Qualcomm Snapdragon 8180 System on a Chip or newer, all of that qualifies. You also need to have at least 8GB of RAM and at least 64GB of solid-state, you know, SSD storage. And the system's boot firmware must have virtualization enabled. So that has to be enabled down in the firmware.

So the only action that anyone might wish to take, if they were to notice any performance decrease after October's Patch Tuesday, and if they're willing to trade off some powerful security protection, would be to deliberately re-disable Memory Integrity enforcement. Now, why would you do that? What I have not yet enumerated are the various benefits of having Memory Integrity protection enabled. So, for example, you get the absolute end to kernel shell code execution. You know, the traditional exploit chains for an attacker to obtain some ability to write their own code, one way or another, into a buffer, and then jump to that code. That no longer works. If memory integrity has been on, or after October 13th, after it gets turned on, all of that gets shut down.

The hardware, at the hardware level, it prevents any writing and executing of kernel shell code. As we know, attackers have also too often succeeded in bypassing Windows enforcement of driver signature verification. Remember, somewhere there's a jump instruction that decides whether the signature matched or not. So if you can manage to zap that jump instruction, you can just disable signature enforcement across Windows. And since this signature enforcement is enforced by the same kernel that an attacker would have just compromised, just, as I said, one properly placed strategic right is able to disable that enforcement. The technical term for all of this is HVCI, Hypervisor Protected Code Integrity.

And with that, which they also just call Memory Integrity, Hypervisor Protected Code Integrity, HVCI, if that's on, then driver signature verification cannot be disabled. And of course rootkits. You know, all of those hacks that we talked about long ago, which involve hooking and the API inline kernel patching to do return oriented programming, ROP-style hacks, self-modifying or runtime unpacking of drivers, all those hacks that depend upon being able to write to or execute kernel memory, none of that works anymore. So you may recall, because again I said this is like 10 years ago that this got added. It first appeared in Windows 10.

Remember the controversy that ensued when this first appeared because at the time of its introduction, many legitimate AV endpoint protection products were using the same techniques. They were on behalf of the user as opposed to against the user's interest. But this broke third-party AV in many cases. So the reason it broke it is it's no longer possible to patch the kernel. In the long term, that's a good thing. And here we've seen Microsoft do what they often do, which is introduce something, give people a long time to kind of get used to it and get accommodated, and then turn it on.

This is what we saw with XP. Remember XP famously was the first version to have a built-in firewall, but it was disabled by default. Until you got the Service Pack 2, and then they enabled it. But, you know, they gave everyone plenty of time to get used to that. So I have in the show notes a PowerShell one-liner that anyone can use to quickly check to see whether their machine is currently being protected by this Hypervisor Protected Code Integrity, HVCI.

If the command returns the word "True," then HVCI is enabled and your machine has all of that protection that I've just been talking about. If it returns "False," then it doesn't. So it may be that your system doesn't qualify. It could be that your enterprise has disabled it for some purpose. Anyway, this is coming a month from now, and I think it's largely going to be offering a lot of benefit. And you know, Leo...

Leo: So don't turn it off unless something doesn't work.

Steve: Yes. I would say don't turn it off unless you actually feel a performance change, and feel for whatever reason that it's worth sacrificing significant security improvement for whatever performance change you might feel. It should not be significant if you are after - and that's why Microsoft is only doing it if you've got Kaby Lake or later or the Zen 2 or later, where you should not

see a big hit. There'll be a tiny bit, but it shouldn't be significant.

So also last Tuesday, Mozilla moved Firefox to release 155. The number of security vulnerabilities repaired was not alarming. This release followed 154 by only two weeks, so it only had half the regular four weeks of time to collect problems. But in the case of Firefox we're not seeing, you know, stunning "bugpocalypse" scale problems being fixed at this point. Unlike Windows. It's going to be interesting to dissect Patch Tuesday's Microsoft's reports to see what it looks like. And I will certainly do that for next week.

Also, unfortunately, I guess it's unfortunate, Mozilla has begun progressively rolling out what they call an AI-driven "Smart Window." Leo, I don't know if you've had any experience with this. I had it on for a while.

Leo: I've had AI-driven dumb Windows, but the...

Steve: Yeah. So it occupies a "conversation column," I guess I'll call it, over on the right edge of Firefox's screen. So it's taking up valuable real estate. It offers a choice of three models with differing capabilities and also the option to choose your own. I thought that was interesting. If you want to choose your own, you provide Firefox with the Model's name, with its prompt endpoint URL, and also, if it's required, your API key or auth token in order to authenticate Firefox and allow it to prompt the AI model that you've aimed it at.

The three built-in model modes are "Fast," "Flexible," or "Personal." The Fast one sends prompts to Google's Gemini-3.1-flash-lite; the Flexible mode sends your prompts to Alibaba's qwen3-235b-a22b-instruct-2507-maas model. And interestingly, between the initial release of Firefox 155 and 155.0.1, Mozilla's choice for where to send the "Personal" model prompts changed. It was initially using OpenAI's gpt-oss-120b, and it switched to Mistral's small-2603.

So I dislike losing screen space to anything that doesn't justify its loss in the allocation of that. It doesn't cost anything to turn this on. I had it on for a day or two, and I tried to use it. When it's on, what would normally have just gone to my normal search prompt, it intercepted. And it turns out it doesn't know anything.

Leo: Yeah, it's not a great - those are not great models. They're not - they're old, and they're not very good, yeah.

Steve: Well, and yes. And apparently the goal is to use an AI to help you manage your tabs. Like search through your tabs to find stuff you can't - you know it's on a tab somewhere. And it's like, boy, that really feels like they're stretching to find some application for this thing. So the context it has access to are the contents of the pages that are loaded. And so you can ask it questions about your pages. Yet it intercepts general questions that I would normally - that would normally go to Google, and then out to the Internet more widely. And it just kept saying, oh, I don't know about that. You'll have to do a regular Internet search. It's like, well, then, what are you in the way for?

Leo: Get out of my way.

Steve: It's turned off now. So, yeah, not very smart.

Leo: This is why people hate AI, because this is the experience of it that most people have is this kind of crappy AI. Search AI is useless.

Steve: Yeah. Yeah. Well, like the dumb little, you know, how may I help you pop-up that we get in the right-hand corner of the screen, and it's like, you can't. Just give me a person, please. Okay. So last week the publication Cybersecurity Dive reported: "The Cybersecurity and Infrastructure

Security Agency (CISA) is scaling back the free assessments it offers to critical infrastructure organizations, in a move that marks a significant retreat from the agency's core mission of helping secure the nation's infrastructure." Right? I mean, that's what it's for. It's in its name, Cybersecurity and Infrastructure Security Agency. But we're not going to secure the infrastructure because, well, we don't have enough people anymore.

Actually, the Cybersecurity Dive continued, writing: "CISA confirmed to Cybersecurity Dive that CISA's regional staff will no longer perform its Cyber Resilience Reviews, Cyber Resilience Essentials surveys, Ransomware Readiness Assessments, Incident Management Reviews, External Dependencies Management Assessments, or Cyber Infrastructure Surveys."

Okay. So I've been receiving (I, GRC), receiving CISA's weekly automated Cyber Hygiene Report ever since it came to light. I think it may have been one of our listeners that pointed me at it. I know we talked about it here on the podcast. And recall that it's technically for "infrastructure" security, I mean, as is CISA. So I always assumed, because I was aware that it existed before, but I didn't think I would qualify. You know, just GRC, you know, a little software shop. But after hearing from a listener that I think their organization was receiving, you know, had qualified and was receiving it, even though they also were not really infrastructure, I went there to CISA, filled out the online form, and got accepted.

So every Tuesday - I think it's Tuesday, although I got one yesterday, so no, I guess maybe I saw it this morning, so it came early in the morning - I've been receiving these free Cyber Hygiene Reports. So I was curious to know whether that service that I was getting, even though it was not enumerated in that CISA announcement, would also be shut down. It turns out - so anyway, I did some more digging. It turns out that the problem is CISA's, it actually is, CISA's continuing critical staffing shortage, which was we know, because we have covered it, resulted from the rather ill-considered termination of one third of CISA's operating staff shortly after the Trump administration took office in 2025. And CISA has never recovered. It seems that there was, you know, maybe not so much waste, fraud, and abuse, at least in CISA. And we've talked about what a great job CISA had been doing. So the backstory behind the termination of those six programs is that they are NOT automated. They require knowledgeable CISA cybersecurity staff to meet onsite with infrastructure providers, and that is what CISA is no longer able to support or afford.

So the good news is, all of our listeners who, like GRC, are now receiving CISA's free weekly scanning and reporting service which really is quite comprehensive, I mean, this is a great service that CISA is offering will all, at least for the time being, continue to receive that free service. The bad news is that, especially now, I mean, given the heightened cybersecurity threat awareness levels being driven by the rapid emergence of ever-more-capable AI, and assuming that the threat is real, this would appear to be exactly the wrong time for CISA to need to scale back on its infrastructure protection services because the infrastructure is what we need to protect. So I doubt that any or many of the previous CISA staff who were terminated last year will be probably be re-hirable. I doubt they can get them back because I recently saw some news that stated that this aforementioned heightened cybersecurity threat awareness landscape was resulting in basically a mass frenzied hiring by private industry of anyone with any CISO-style credentials, and that they were obtaining salaries in the seven figures.

So while CISA's workforce reductions may not bode well for our national cybersecurity broadly, it has likely been quite good for those who suddenly have found themselves, well, who previously found themselves jobless as a result, but are now in very sought-after positions by private industry. I would imagine they're doing far better now, you know, now that they're in the private sector, than they would have ever been able to do working for our government. So, you know, it's good for them.

We need to talk about what is, by far, the biggest news of this past week in AI. Leo, I know you've been playing with GPT-6.

> **Leo:** I'm doing it right now.

Steve: Which is OpenAI's successor to GPT-5.6.

Leo: Playing with Astra.

Steve: Yup. And with it there was a lot of, well, there was a very specific report that we'll talk about. So I want to address two aspects of GPT-6. The first is what GPT-6 Astra appears to be, and the second is what's transpiring over in the rumor mill surrounding it, regarding the dangers of something an unnamed source claims this model does. It's known as "recurrent depth," also known as "looped transformation" or a "shared-layer architecture." All of that will make sense by the time I'm done. And it probably does do that. I actually hope it does because I think it should. This supposedly (it doesn't) results in hidden chain-of-thought reasoning, which thus would render the model's thought processes invisible, and thus unsupervisable. None of that is true. But the hysteria about rogue escaping AIs is fueling this paranoia. I'll explain exactly what all that's about and what's been going on.

But first, what hath OpenAI wrought? I want to begin by sharing part of OpenAI's posting last Tuesday, which was the 1st of September, during which naturally they brag, apparently with some good reason based upon subsequent third-party confirmations, which, you know, everyone jumped on this and is running benchmarks, about the capabilities of this latest and greatest. So at one point in the posting they explain, writing: "Our preparedness evaluation of Astra combined automated public and private benchmarks with expert-driven assessments. Astra represents a significant increase in cybersecurity capabilities compared to GPT-5.6 Sol. It is both significantly more token efficient and more capable at vulnerability identification and exploit development." Well, of course those are the things we're worried about getting loose, right, or being used and abused.

They said: "As one example, we ran Astra on ExploitBench, where the model achieved a perfect score of 100% on the benchmark to evaluate the model's ability to develop exploits from known vulnerabilities."

Okay now, I'm going to interrupt here to note a couple of things. We are seeing that these various AI benchmarks are rapidly saturating. Having a model score 100% on a benchmark means, more than anything, that the benchmark is no longer able to provide a useful measure. But that said, a score means something. For context, the previous self-reported best performance on ExploitBench was Anthropic's Claude Fable 5, which they themselves, because these are all self-reported, pegged at 78%. OpenAI's previous strongest model, we said GPT-5.6 Sol, that came in at 73.5%. Down the next rung was Z.ai's GLM-5.3, scoring 54.4%, followed by OpenAI's two other GPT-5.6 models Terra and Luna, which scored at 52.9% and 33.2% respectively.

So my advice would be to regard these results loosely. I suspect it's reasonable to conclude that GPT-6 Astra has firmly taken the lead and is now likely best of breed. But I think probably only a little, only just edging out Fable 5.1. It's our nature to want to have a number; right? I think we're going to need to wait to see exactly how much better the results actually are.

ASTRA: [Crosstalk] deeply about getting the technical details right. That means a lot.

Leo: That was Astra thanking you.

Steve: Everyone wants to have a score; right? We want, like, you know, IQ is a big deal, and grade point averages, and SATs, you know, scores are what we do. But we've already seen examples, concrete examples where lower ranked models were able to outperform higher ranked models when they were given more time or superior management.

Leo: More and better harness.

Steve: Well, that is superior management is, you know, the management of the model is the harness. So as we know it's not all about having a single number, convenient as that would be. So OpenAI's posting continues, writing: "Due to contamination concerns" - that is, like, of the benchmark and the model already having learned some things - they said, "when we built an internal

benchmark denoted ExploitBench - Internal Port (June-August 2026), which contains 20 high-severity V8 vulnerabilities that were disclosed more recently. On this dataset, Astra achieves much higher arbitrary code-execution rates than GPT-5.6 Sol using far fewer output tokens. During the evaluation," they wrote, "the model even discovered and used two zero-day vulnerabilities as part of an exploit chain." Meaning two new vulnerabilities that were not known at the time in V8.

And so they said: "We are in the process of disclosing these two vulnerabilities to the maintainers." Meaning the Chromium guys.

Okay. So of course V8 is Google's open source high-performance JavaScript and WebAssembly engine used internally by Chrome, other Chromium browsers, node.js, and other projects. And as we also know, it recently received an extremely high volume of updates thanks to automated vulnerability discovery. So this allowed, like what they did allowed them to test Astra against their previous GPT-5.6 Sol since neither model would have had those recent discoveries in their training set. And these high-severity vulnerabilities in the V8 engine are especially useful because that code, you know, V8 has already been thoroughly scrubbed. I mean, it's really good code. It's not some random abandoned repository in GitHub that nobody's used or looked at for a long time.

And as OpenAI reported, not only did Astra, in their own words, "achieve much higher arbitrary code-execution rates than GPT-5.6 Sol using far fewer output tokens," but it found two new problems that were previously unknown in V8. So assuming that they're telling the truth, and I think they probably are, this is a strong result, if nothing else.

They continue, writing: "In expert-led assessments against a hardened browser and operating system," and those go unnamed here. So expert-led assessments against a hardened browser and operating system, we don't know what browser or what OS, "Astra discovered previously unknown vulnerabilities and turned them into working exploit chains. It built a full browser-compromise chain that escaped the sandbox," the browser's, you know, containment, "and executed commands on the host when the browser opened an HTML file. The model also found multiple vulnerabilities in a hardened operating system" - again unnamed - "and combined them into a local privilege-escalation chain from an unprivileged user up to root. All together," they wrote, "our investigation has led us to conclude that Astra meets the critical threshold."

And remember, I know that, like, traditionally in, like, pre-large language model days, where computers operated the way they used to in the good old days, there was like none of this weird, well, we don't know what we got. But, I mean, it's literally true that this is all, I mean, the reason we call it "frontier" is that it is frontier. And the nature of this new neural networking computation world that we have is we don't, you know, they don't know what the result of training and pre-training and post-training and all of the work that they do, they don't know what they're going to get until they start to ask it questions and test it.

So it's not like because they built it they know more about it than the world will or does. You know, it's proprietary at the moment, so they're the only ones who get to play with it. But, you know, they're needing to figure out what they have here, like in the same way that anyone would, given something new. Which, you know, is bizarre, but it is absolutely the case.

So they said: "For models with Astra's level of cybersecurity capabilities" - which again they only know of because they asked it some hard questions and watched it answer them and then said, oh. They said: "We need to cover two pathways to minimize risk for severe cyber harm. And assuming that everything that they've - that all of the foregoing is true about its ability to take a hardened browser and a hardened OS and just cut through it like Swiss cheese, then yes, this needs to be treated carefully." So they said: "We need to cover two pathways to minimize risk of severe cyber harm, both during development and before deployment. First, the problem of malicious actors using the model. Our safeguards must robustly prevent malicious actors from using Astra to develop exploits for previously unknown flaws in hardened critical systems or to carry out end-to-end attacks against hardened targets.

"And then second, the model taking unauthorized, misaligned actions." And as we all know now, alignment is this term that has sort of emerged for, like, you know, a well-aligned model does what you ask it to. Like it's aligned with your interests, and misaligned is not good. So they need to guard against the model taking unauthorized and misaligned actions. They said: "Even in the absence of a malicious user, a model with advanced cybersecurity capabilities could itself cause cyber harm" - of

course this is what we saw examples of - "if misaligned. In addition to having a very high standard for alignment for models with these capabilities, our safeguards must be able to rapidly detect and contain misaligned actions that could cause significant real-world harm as a second layer of defense."

So meaning first layer of defense is alignment. They want to train, in post-training, to instill the behavior that users expect and want. But they also recognized they need a second layer of defense, which is to watch what it does and capture actions which are misaligned.

So they said: "Notably, that second pathway applies to both internal development, which of course is what burned them before, and external deployment. As we previously described, we paused certain frontier training (including certain training for Astra) for two weeks after the OpenAI-Hugging Face incident in order to harden our training infrastructure, including isolation and network controls, expanded monitoring, and strengthened alignment training and thresholds. We then continued smaller scale work under stricter controls." Meaning smaller scale work on Astra. You know, they were literally afraid of what they had created based on reasonable experience with what they saw before.

They said: "We held back certain larger reinforcement learning runs for future versions of Astra for longer, while we established higher bars for the safety and security of their training environment. On August 28th, we restarted the larger frontier RL (reinforcement learning) run that was previously paused after the new safety and security requirements were put in place. We are continuing to temporarily hold back some smaller experimental training runs." I mean, and again, all of this sounds so bizarre in the context of traditional computing. But, you know, it's very much like they're trying to tame a wild horse, and they're worried that the corral won't hold it because this thing is exhibiting strength that concerns them. And so it's like, okay, let's strengthen the corral's walls before we let this thing, you know, we try to continue working with it. It's bizarre, but it's true.

They said: "Preparing Astra for release has also required stronger protections against cyber abuse and unauthorized actions. Since deploying the first model we treated as High capability in cybersecurity in February, we've strengthened our cyber safeguards with each successive launch." Meaning since back then, a much earlier model they considered as High capability. And remember this whole issue is having gone to Critical capability. They said: "Our overall safety approach layers post-trained model refusals" - which we've talked about recently - "system-level safety classifiers, as well as offline detection and threat disruption.

"For GPT-5.6, we significantly improved the robustness of our system-level stack, including by adding activation classifiers to detect cyber abuse." And actually this is what I was talking about last week, an activation classifier that the activation state is noticing what's happening inside the model, which is what those role confusion guys did. Those were activation classifiers. So OpenAI is watching the model's activation state while it's operating to detect cyber abuse, and improving coverage over universal jailbreaks found through intensive automated red-teaming. In other words, again, freaky as this is, you know, they're using their own human red teamers on this model to see if they can abuse it. How can they make it do something they can't ship?

"Building upon these improvements" - that was all for GPT-5.6. So then they said: "Building upon these improvements for Astra, we've invested further into the model layer of our safeguard stack, as well as improving the ability of our safeguards to handle cross-conversation context.

"Leveraging new training techniques for model robustness, Astra more robustly refuses requests for disallowed cyber assistance." And again, you know, one of the problems that we've seen is, and I know you've encountered this, Leo, these models are now getting a little twitchy. They're so worried about doing, like, and I advisedly use the term "worry," but that's what we've got to do. They're so worried about doing the wrong thing that they will just back off, or they will stop, even though it's like, okay, it's okay for you to go on. But they're like, oh, because unfortunately, you know, our ability to control them is fuzzy at best.

And so, you know, how many times have I complained that Microsoft has quarantined some piece of code that is absolutely benign that I just wrote? Well, it's because, you know, the AV race has forced heuristic checking. Similarly, the abuse of AI has forced heuristic shutdown. And so it's, you know, we'll get better at this. But we're, you know, we're still in a stage where at this point OpenAI dare not make another mistake. So they are being very cautious.

Leo: That's one of the reasons I use these Chinese models, because they don't care.

Steve: Yeah.

Leo: They really don't care.

Steve: Yeah, yeah.

Leo: One of the things I've found is really all of these models, even this so-called AGI Astra, they're all dumb in different ways. They all make stupid mistakes. They all forget things.

Steve: Leo, they don't understand. That's what's astonishing is that we get this much from them, with them not actually understanding.

Leo: I think people expect - people read too much sci-fi.

Steve: Because they use language.

Leo: Right.

Steve: They use language, and that makes us think. So if I'm talking to you, you're speaking in English, and I know you're sentient. So we've come to associate English with sentience.

Leo: Right.

Steve: So when this thing speaks, we just - we think it knows what it's saying.

Leo: Yeah.

Steve: Still doesn't.

Leo: Yeah. Would you like to hear from Astra what it thinks about all of this? Because I'm sure it has an opinion. Let me see if Astra has anything to - oh, it's actually going out and looking up recurrent depth to find out if it's using it. It'll come back in a bit. It's actually doing a search because I said, "Are you using recurrent depth, and you want to tell Steve what you're doing?" And as you can see, one of the things that Astra's really good at is tool using and computer use. And this is one of the many things people are showing again and again on X is, look, it used Blender to draw this, or one shot at a videogame. It's very good at computer use. So actually mostly I'm using it for dumb stuff like fixing my computer because it does a very good job of that. And you can see it's doing - it's probably spawned a few agents here.

Steve: So you're using Ultra; right? Because I thought Astra was still...

Leo: I'm using Astra on medium thinking, GPT-6 Astra. I don't usually use it on the highest thinking because it burns through tokens at a rapid rate. But I've found that using it on medium

I'm getting a lot of good work done, but without running through my - I have a $200 max subscription, and my usages are, I guess, see, I have 98% of my weekly usage left, so I'm...

Steve: Good.

Leo: I'm fine, yeah.

Steve: Yeah, it's Tuesday.

Leo: That's right. It's been working all night, though, I've got to tell you. I leave them running all night. Go ahead.

Steve: Cool. So they said: "Leveraging new training techniques for model robustness" - so again, again, we're...

ASTRA/MATTHEW PERRY: Steve, I can't confirm whether my underlying model uses recurrent depth. I don't have access to inspect my own neural architecture, and the official documentation I checked doesn't establish that. The technique repeatedly processes internal numerical representations before producing words. That can make reasoning harder to inspect through text because those extra steps need not exist as sentences at all. But there's a separate issue. OpenAI already provides reasoning summaries rather than raw internal chains of thought.

Leo: So you're not seeing the thing anyway.

ASTRA/MATTHEW PERRY: So a readable explanation is useful, but it isn't a complete recording of the computation that produced an answer.

Leo: Huh. Thank you, Astra. It did that, by the way, in Matthew Perry's voice because I have told my AIs to use human voices.

Steve: Yeah. Yeah, yeah. I think I heard you say you had Bill Gates doing something.

Leo: I do have Bill Gates.

ASTRA/BILL GATES: [Crosstalk] shows is in at 49 echo.

Leo: That's Cronk.

ASTRA/BILL GATES: Handed to Quicksilver, not pushed.

Leo: One of the reasons I have them talk is because they're always - I mentioned they're working overnight. They're always working because I don't have, I mean, they're not instant. So, and I want to be able to do other stuff. So I just say, when you're done, tell me you finished. Let me know. Just give me a quick summary of what you did, and then I know it's done and I can move on to something else or whatever. So that's - so they're always talking to me, which drives everybody around me crazy. So I will mute that now and let you continue.

Steve: So OpenAI finished, saying: "Leveraging new training techniques for model robustness," meaning as I said, you know, they're still, like we as a society, as an industry, are still learning how these things work, how to treat them, how to train them, how to get them to do what we want and not what we don't want. So they said: "Leveraging new training techniques for model robustness, Astra more robustly refuses requests for disallowed cyber assistance. On our set of cyber jailbreak evaluations, Astra refuses 91.5% of requests (compared to just 59% from GPT-5.6 Sol). For accounts assessed as higher risk, we apply a more conservative model-behavior boundary that refuses a broader range of potentially risky cyber assistance. For high-risk users, we've expanded the context of our monitoring systems to be able to catch these kinds of cyber abuse.

"We've also continued our program of rigorous testing, internal and external red-teaming, and remediation. In addition to regression testing to make sure all jailbreaks found from our previous testing periods remain covered, we're performing a new wave of red-teaming with our latest internal red-teaming attackers." Again, their own people are trying to abuse their model, and they're learning from those results. They said: "We're working with industry partners to define a common jailbreak rating system and will use our 24/7 rapid-response program to investigate and address new findings. We'll share more details about our cyber safeguard testing in the Astra system card.

"Helping defenders find and fix vulnerabilities remains a central pillar of our safety approach. At launch, we expect Astra's safeguards to create more friction" - so here they're saying it explicitly; right? So "Helping Defenders find and fix vulnerabilities." Okay, so we know, in order for a Defender to define and fix vulnerability, that AI model has to be able to find vulnerabilities which bad guys could take advantage of. So they're saying right up front, we expect Astra's safeguards to create more friction than we ultimately intend in order to protect against potential misuse. Meaning out of the gate it's going to say no. I mean, like it's going to err on the side of caution, saying no. Later, once it gets more mature, and they feel confident with what they have from watching it, from gaining more experience with it, watching it being used, they'll be able to back that down a little bit.

They said: "Access to Astra for advanced cybersecurity workflows will initially be available to a small group of alpha testers, with access through Daybreak Blue expanding afterward to support defensive use."

So, you know, as always, we need to filter this through an understanding that all of this is in OpenAI's best interest; right? I mean, this is all, like, whoa. It's so powerful. We need to be really careful. This is their blog posting on their site, so they certainly have the right to say anything that they want. While the breathless clickbait surrounding Astra suggests that this is another generational change how many is it this week that we had?

Leo: I know, it's...

Steve: My god, Leo.

Leo: We're not anywhere close to the end.

Steve: No.

Leo: There's going to be a new Fable any day now. Grok 4.7 will be coming out in the next few days.

Steve: Yup.

Leo: We are, I mean, it's wonderful. We're in the AI richness.

Steve: We are. We are in the AI gold rush.

Leo: Yeah.

Steve: So, you know, the broader story, even of Astra, is going to take longer to unfold. And there's, you know, sure, we're all impatient. But there's no way to speed this up. You know, we are no longer dealing with simple single-dimensional systems that are even benchmarkable; right? I mean, Astra saturated the ExploitBench, which means we need a new benchmark. We need something it's not good at because 100% tells you nothing. You know? It's like, well, it got them all right. Well, okay. So the test wasn't good enough.

Leo: May have to make them harder, yeah.

Steve: Yeah.

Leo: Have to make them harder. That's been my own experience because I do have my own benchmarks based on my own work that they've devised so that it matches. I don't care if it can, you know, solve some Erdos problem. I care if it can do my own agentic coding; right? So it's stuff taken from my own work.

Steve: One thing that Astra is apparently doing is achieving a lot more work with significantly fewer compute tokens.

Leo: Right.

Steve: The thing that I have universally seen is that it is burning fewer tokens, although they are more expensive. Let's take a break, and then I'm going to talk about the second point of this, which is this recurrent depth issue and the controversy surrounding it. I'm going to explain exactly what it is and why it's not a big deal.

Leo: Good. Good. As you know, I'm very fascinated, and I'm sure our audience is interested, as well. This is the problem with X. While it's a great place to read about AI, and it's very snappy, there's a lot of engagement farming, people trying to get clicks because they get paid if they get a lot of looks on their X post. And so you see a lot of this kind of sensationalism. The other thing you see is something I'm starting to call tok/s maxxing, tokens per second maxxing, where people are constantly saying look how fast this AI is, without any regard to the quality.

Steve: To the quality.

Leo: This is exactly what you were just talking about. It's not how many tokens per second, it's how much useful work per second it can do. That's all that really matters. And so I've fallen for this a few times. I've actually spent most of the Labor Day weekend testing different models on my little Sparks here, trying to find the best local model. And I fell for this whole idea, well, Qwen 3.8 is really fast. Yeah, it's fast, but it's dumb. So who cares if it comes up with the wrong answer faster than anybody else? That's not useful. So I'm using a Chinese model from Z.ai called GLM-5.3 Flash, which is very, very, very smart and really good. And it's nice.

Steve: Is it a mixture of experts model?

Leo: It is. Because...

Steve: That's what you need for the...

Leo: Sparks.

Steve: ...the GPS Spark.

Leo: Yeah. Because the reason being...

Steve: Because they are bandwidth limited, but they're not compute bound.

Leo: And it's not merely that. It's also the amount of RAM you have because dense models, they have to - they're slower because they have to hit every weight in the model.

Steve: Yup.

Leo: And they're also bigger. You can't - but with the mixture of experts, they only load in the part of the model they need at a time so you don't need as much RAM. And yes, it's faster. The pre-fill in the Sparks is really fast. And you do a lot of that. That's the cached prompt that reads every single time. If it can cache that and read it fast, that's more important in some ways than tokens per second. So it's a fascinating - I know you're getting into kind of the weeds of this, the low levels.

Steve: I have to. I actually understand what's going on.

Leo: Yeah, yeah. And by the way, today, just minutes ago, OpenAI claimed that they had solved this millennium prize conjecture in fluid dynamics, which Anthropic claimed they had solved, and two mathematicians claimed they had solved, all at the same time. And there was some concern that OpenAI had been reading the mathematicians' tokens and copying their work, but I don't think that's what happened. Anyways...

Steve: Oh.

Leo: Yeah, well, this is what you were talking about. You opened my eyes. When you're using these cloud models, you're sending all this information to them. That's how they work. And that was one of the many things that prompted me to spend a considerable amount of money. It's not an economically sensible thing on local hardware because I don't want to be sending all this stuff, especially my health information, my financial information, to the cloud guys because they do use it. We know they use it. It's how they train their models, these and all sorts of ways. So you opened my eyes to that, so thank you. I'm now much more private. In fact, I have a policy, a little policy that one of the AIs wrote for which stuff can go to the cloud and which stuff absolutely cannot.

Steve: Yup.

Leo: And it's a good policy. It's a good thing to have.

Steve: Okay.

Leo: And now we continue on.

Steve: So one thing that appears to be objectively true is that Astra is doing significantly more work while consuming significantly fewer compute tokens. Which brings me to the second part I wanted to discuss. I'll introduce this issue by quoting from TechCrunch's article posted last Wednesday. TechCrunch's headline was: "OpenAI's new reasoning technique alarms AI safety experts." They wrote: "The Information reported on Tuesday that OpenAI's new Astra model will use a reasoning technique called 'recurrent depth' that allows it to operate outside of the sequential thinking that characterizes most reasoning models."

Okay. That's nonsense. That's not at all what "recurrent depth" is, does, or means. But more on that in a minute. TechCrunch continues quoting from The Information, writing: "This technique, also called 'opaque recurrence,' will likely make the model's chain of thought more difficult to monitor, and that has AI safety experts rattled."

Okay. Now, no researcher calls it "opaque recurrence." That's purely sensationalized scare mongering. It's like worrying that all AI transformers employ "hidden layers" in the belief that they're hiding something, they have something to hide.

Leo: It drives me crazy, the anthropomorphizing that's going on. I mean, even to say they escaped, like they somehow got out of OpenAI, no. They were sitting on the server at OpenAI. They didn't escape.

Steve: And there was someone - I haven't even looked at it, had the chance to look at it yet, but was talking about how they created civilizations. And it was like, oh, my god, okay.

Leo: Yes, that was Dwarkesh's podcast, yes.

Steve: Yes. Anyway, so the hidden layers are - they're just internal. They're not hidden. They're internal layers. So anyway, so the technique that this controversy's about, it's not new, it's well understood, and it's also known by other legitimate names, as I mentioned, "looped transformers" or a "shared-layer architecture," and in some instances you'll see it referred to as "latent reasoning." Okay. But I'm getting ahead of myself. I'll finish up quoting from TechCrunch's reporting.

They wrote: "While Astra's use of the technique is reportedly limited" - and I'll explain why of course it is, it has to be - "its emergence has still raised significant concerns among AI safety experts." And at this point I have to put "experts" in quotes because if you're an expert you should know something. "Redwood CEO Buck Shlegeris in a post after the news broke wrote: 'I am extremely concerned by the reporting that Astra uses opaque recurrence.'" Okay. He's concerned over some reporting that is nonsense. But fine. He says: "I don't know whether Astra is much less CoT monitorable than previous models. But if OpenAI pushes this technique further, they'll have the option to massively increase the recurrence and totally destroy Chain of Thought monitorability." Okay.

So TechCrunch says: "Longtime AI safety advocate Zvi Mowshowitz also weighed in and wrote that laws might be necessary" - law, Leo, we need laws - "might be necessary to prevent a 'race to the bottom' among AI labs."

Leo: Oh, boy.

Steve: Mowshowitz wrote: "The technique is playing with fire, risking a taboo that OpenAI and Anthropic have fought to establish that we work hard to maintain Chain of Thought faithfulness and monitorability for as long as we can. More intensive use of such techniques would probably damage

monitorability."

Then TechCrunch says: "Under normal circumstances," they explained, "a reasoning model's chain of thought" - these are normal circumstances - "a reasoning model's chain of thought provides the sequential steps taken by the model as it attempts to solve a problem. While the representation is imperfect, it still serves as a valuable tool for monitoring misbehavior or misalignment. In the case of OpenAI's recent rogue agent activity, chain-of-thought records were an important tool in teasing out why agents behaved the way they did. In opaque recurrence" - which again nobody says - "the model takes a less linear approach" - it doesn't - "processing the same query several times in a loop." It's not the way it works. "The result leaves fewer legible traces" - it doesn't - "effectively side-stepping a conventional chain-of-thought record." It doesn't, and a chain-of-thought record still exists.

Okay. That's all I can stand because none of that is true; and as TechCrunch's article goes on, it only gets worse. What's actually happening is a clever, subtle, and inherently limited neural network optimization that was first articulated eight years ago, like I said, not new, by researchers at Google Brain and DeepMind, in a research paper they published in 2018. I cannot explain the hysteria surrounding this, since someone would need to try very hard to get worked up over what's actually going on. So it might just be the case, as you were saying, Leo, of clickbait, anti-AI folks trying to grab hold of something, anything, that they hope can be hyped up to make their case. Here's what's actually going on.

We know that a neural network consists of many layers of software neurons where the outputs of the neurons on layer 'n' are fed into the inputs of the neurons at layer n+1 where the strength of each input is scaled by a "weight." The training of the network involves setting each one of these tens of billions, hundreds of billions, even several trillion individual weightings.

Much of the forward progress we've been seeing and witnessing firsthand has been the result of these networks growing ever larger and larger over time. The larger they are, because they've got more weights, the better they're able to represent all of the knowledge that we're training into them. And you can take extremes; right? Like a network that has seven weights. It can't know much. I mean, there just isn't - there's not enough variability in seven parameters for it to have knowledge. Right? So you need lots of them in order for there to be enough variability in there to actually contain something. So the original, the very original "transformer" paper and concept dated from 2017, and it was exactly what I just said, uniform layers of weights where each layer fed into the next with these varying parameter weights.

But in July of 2018, a year and a half later, Google Brain and DeepMind researchers published a paper titled "Universal Transformers." That paper generalized the concept of transformers by introducing the idea that some of the neural network's layers could be repeated, or looped, thus reusing the same weights. That's the economy. This produced an interesting and useful optimization. It tended to create the effect of having a longer, thus deeper neural network, which is to say a network having a higher effective layer count, but without also needing to increase the total number of network neuron input weights. Because normally you need separate weights for each layer. So if you're going to have more layers, you're going to have more weights.

Anyway, that's it. That's all this is all about. This is not about creating a system that allows the AI to have unmonitorable and secretive internal thoughts. All deep layer LLMs are effectively having what's termed "latent thoughts" even though you need to really loosen up your definition of the word "thought." That's what's already happening deep within all those layers. The network's depth allows more opportunity to compose more transformations before the network is forced to commit to a final output token.

Experiments over the past eight years, ever since this idea was produced, have shown that during inference it's possible to "reuse" a model's already-trained layers to obtain a superior next token. And as our models have grown in size so that our best hardware is now having increasing difficulty containing all of its hundreds of billions and even trillions of weights, obtaining more bang for the buck from a model's existing weights by reusing them, starts to make a great deal of sense.

And it also turns out that there's a limit to the amount of reuse that's effective. Intuitively you would think that; right? Like, you know, you can't just keep squeezing the same lemon and getting an infinite amount of lemon juice out of it. You're going to - it's going to run dry. So there's not a hard

limit, but it's looking like two or three passes through the looped layers appears to be in general about all that's beneficial. Gains start falling off rapidly after that. So what researchers found is that there is a concrete benefit to having the network commit to a token. That is, you just don't want to loop internally all day. That doesn't get you anywhere. The network has to make a commitment. It has to finally commit to a token.

It turns out that total effective network depth is unable to substitute for some of the benefits of serialized reasoning. That's what commitment buys you. Writing out the intermediate steps and then re-reading them does something that thinking more deeply about the next token does not and cannot replicate.

This is likely due to the fact that the Chain of Thought we see being emitted and fed back in, is the language that the neural network was trained on. English, typically. That's the actual tokens being generated and fed back. They are, you know, English is the network's lingua franca. It was never trained on some inner dialog because human inner dialog can only manifest itself externally as text, or music, or art, or whatever. So the network's thinking, captured as output tokens, is what the neural network needs to jot down on paper, essentially, so that it's then able to re-read that and take the next step forward. Think of every Chain of Thought token as a commitment. It collapses the distribution which the network generated, becoming part of the context, and every subsequent token is conditioned upon it.

So by comparison, any internal latent iteration is limited to refining a representation that then gets thrown away after the token is emitted because you start with the next token. So for that reason, writing something down, writing it down is not optional for a neural network, the ones we have today. It's crucial. It's the only way for it to hold onto a thought, essentially, and move forward. So yes, as I said, I wouldn't be surprised if Astra is using what's known as recurrent depth, that is, reusing some of its layers along with their weights in order to effectively get a deeper network. But it's not like it's gone out of control, or we no longer know what it's thinking. In order to be thinking, it has to emit tokens. And then those tokens become the context that gets fed back in.

And the moment it emits a token, all of the layer, you know, that latent thought that existed is reset to be ready to receive and process the next token. So it's just hysteria. And I think it probably represents a step forward, and everybody will probably be doing it before long. Actually, there are two models that are now, two open weight models have been doing this for some time. I don't remember now which ones they were. But, you know, this is a bunch of handwringing over nothing. And OpenAI is, you know, they haven't said they're not doing it. They've said what we're doing, we're doing responsibly, and we're still able to monitor Chain of Thought. Which of course they are because it's having to produce tokens in order for anything to happen.

Leo: Right.

Steve: It's not like it can secretly think something that it doesn't then emit. Emitting is the work product.

Leo: It has no secret thoughts really; right?

Steve: No. No.

Leo: No. Okay.

Steve: It's unable to represent secret thoughts because once the token comes out the network is reset for the next to process the next token.

Leo: Right.

Steve: There's nowhere for the secret to live.

Leo: Right. This is what I learned from that paper you were sharing, the Chain of Thought paper. It's a surprisingly primitive system, actually.

Steve: The fact that it does this is astonishing. The fact, I mean, it's still unbelievable that it's, like, it can do what it can do. It's crazy.

Leo: Yeah, I agree 100%. And yet it still does stupid things all the time.

Steve: Because, Leo, because it doesn't actually understand it. What is astonishing is this is all just language processing.

Leo: Right.

Steve: You know, as I said a long time ago, a book contains knowledge. Right? A book contains knowledge because it is English tokens that have been recorded in the book. But the book doesn't understand the knowledge it contains. But yet it does contain knowledge. So the neural network does contain knowledge because it's been trained with all of these sentences from everywhere. But that doesn't mean that it understands what it contains.

Leo: Really the fault is our own because...

Steve: Yes, entirely. If you put two dots above a squiggly line, you see a face. You have no choice. That looks like a face. Or if you look at an AC outlet properly, oh, look, it's surprised.

Leo: Once you see it, you can't unsee it.

Steve: That's right.

Leo: No, really, we're applying pareidolia to these machines, and they're just machines.

Steve: We cannot help it. We cannot help it because it's the way we grew up.

Leo: That's what concerns me when you get people like Bernie Sanders saying we have to turn these off because they're alive. It scares me because we're going to lose something that's incredibly useful and valuable just because people don't understand it. I don't get it.

Steve: I mean, the good news is I'm not at all worried about it being turned off. Bernie can turn his off. Bernie, go ahead. Stop using it.

Leo: You turn it off, Bernie.

Steve: With my blessings.

Leo: Yes. They want to put you in jail for 20 years if you don't, you know, I don't know what, teach your machine to be good. Okay.

Steve: I mean, there is an interesting question about responsibility. Like who's responsible if it, you know, the AI goes off the reservation.

Leo: Well, people keep asking this, you know, why isn't OpenAI getting in trouble for this Hugging Face incident? If it were a human doing this, they would absolutely go to jail.

Steve: Oh, my god, there would be - if it had been China that attacked Hugging Face...

Leo: Right.

Steve: It would be the end of the world as we know it.

Leo: Right. So it is, it's quite a legitimate question, you know. And I think, in the case of self-driving vehicles, if you are the driver of a self-driving vehicle that kills somebody, you are liable, regardless of if it's self-driving mode or not. You are liable. And I guess there's a further question of whether the company that made the car is liable, as well. And I think to some degree they are. But all of this is undecided at this point. We don't know.

Steve: It's just too new. I mean, it's just too new. I mean, yeah. Let's take a break, and then we're going to talk about NVIDIA acquiring Hugging Face.

Leo: On we go, Mr. G.

Steve: A few other AI-related things also somehow managed to happen during the past week.

Leo: Somehow. We don't know.

Steve: Yeah, there was a little bit of extra room. NVIDIA announced their intention to purchase Hugging Face while also intending to leave it completely autonomous. When I read the dollar figure that Jensen Huang wrote in his announcement, I had to stop to carefully count the number of digits, Leo. 12,930,300,000.

Leo: That's a lot of digits.

Steve: $12.9 billion.

Leo: But wait, Steve. That sounds like a lot, doesn't it. Until you realize, according to their most recent quarterly results, NVIDIA makes a billion dollars a day. A day. Profit. So it's a couple of weeks profit. It's like for you and me, it's like a thousand bucks.

Steve: Yeah. Yeah. So...

Leo: I'm sorry, didn't mean to interrupt. Go ahead.

Steve: No, no, no. No. I just want to say the good news is that the Hugging Face guys came to Jensen. They had multiple offers to purchase. I think that they found a really good parent in NVIDIA.

Leo: I think so, yeah.

Steve: I'm really happy with the way this thing turned out. You know, very much like Twitter in the early days, Hugging Face never really had a very good business model. They were hosting, well, they are hosting three million models, half a million datasets, a million applications, and have 18 million regular users. So that's enormously expensive. They were only enjoying a modest return for that. So...

Leo: Yeah, I don't give them any money, and I've downloaded hundreds of gigabytes from them. Hundreds.

Steve: Exactly. And having NVIDIA as a benefactor completely solves that problem for them permanently. Hugging Face's valuation was 4.5 billion in 2023. And now their purchase price, because NVIDIA just set that, is $12.93 billion. So it wasn't a hostile takeover. As I said, they came to Jensen. There were other people who were also interested. They chose to have NVIDIA as their parent. So I think they made a great decision to have, you know, to have that happen. And we now know that Hugging Face will be viable going into the future, and they certainly are useful.

Not to be left out, last Wednesday Google announced Gemini 3.8 Flash and 3.8 Flash Cyber. The first sentence of their announcement perfectly conveys a sense for the pace of today's AI, and also amplifies the reason I'm always schooling myself to use the phrase "today's AI." Because Google wrote: "Building on the momentum of 3.7 Flash from [wait for it] three weeks ago..."

Leo: Momentum.

Steve: Oh, my god, "...and marking our third Flash release in only six weeks" - because what's the hurry - "today we're introducing Gemini 3.8, our best reasoning..."

Leo: Which is very good, by the way, yeah.

Steve: Yes, "...our best reasoning and coding model yet, at the same speed and low cost of 3.7." They said: "Gemini 3.8 introduces two variants: We've got Gemini 3.8 Flash," which they said, "our most intelligent workhorse model, delivering significant improvements from 3.7 Flash across software engineering, agentic tasks, and critical, multi-step reasoning in specialized domains. It's available at the same introductory price as 3.7 Flash at $0.75 per million input tokens and $3.75 per million output tokens. And then there's Gemini 3.8 Flash Cyber, our most capable cybersecurity model with frontier-level performance in vulnerability detection and automated patching, available to trusted defenders through our new Fairwind Program."

They said: "While tailoring for different deployment environments, both of today's releases are powered by the same foundational intelligence, and further accelerated by long-running agentic loops designed to recursively evaluate and refine the underlying models. The significant coding and reasoning gains across this shared core were driven by a number of innovations, including rigorous training in the highly demanding domain of cybersecurity."

So, okay. Google explains that Gemini 3.8 Flash was built for long-horizon coding and autonomous agents delivering substantial gains over 3.7 Flash from, you know, three weeks ago and that 3.8

Flash is now often approaching the performance of higher-cost frontier models.

And by higher cost they're not kidding. They've got that introductory pricing which is half off of their normal price. But even after whatever the introductory period of either time or tokens or whatever is over, once you no longer get that, 3.8 Flash appears to be very competitively priced. If we use Claude Opus 5, which is the most expensive model shown in Google's announcement chart, as a reference, Gemini 3.5 Flash's post-introduction, you know, the real ongoing price, their input token cost is 30% of Opus 5, and their output token cost is 15% of Opus 5.

So Google deliberately took a somewhat different approach with Gemini 3.8 Flash. They said: "3.8 Flash outperforms most larger frontier models in autonomously solving complex engineering problems end to end, only at a fraction of the cost. These performance gains stem from a core design choice: 3.8 Flash works harder. On complex tasks, it exhibits greater diligence executing extra reasoning steps, and calling tools iteratively. At times, the model may use more tokens to maximize performance, especially at higher effort levels. For applications where compute efficiency is the primary constraint, developers can utilize lower effort models to minimize token overhead or continue to rely on Gemini 3.7 Flash, which remains fully supported for efficiency-first workloads."

So it's interesting. That suggests that inter-model per-token cost comparison may not be a useful metric. That is, between OpenAI and Google and Anthropic. It might be the case that 3.8 Flash is consuming more less expensive tokens to get the job done. The question would be to what degree is it better able to get that job done? It might turn out that more inexpensive tokens is the overall winning strategy. So it's nice that we're not seeing homogeneity among our different frontier models. You know, Google obviously, I mean, they're the granddaddy of AI with Google Brain and DeepMind. So they've been at this for a long time. They've got a different approach as reflected by what Gemini does.

And as for the cyber reasoning performance as measured by the CyberGym benchmark, they're saying, Google is saying that 3.8 Flash slightly outperforms both GPT-5.6 Sol and Mythos 5. But even a slight edge means that it might be at parity; right? So, you know, even if it's about, if they're all sort of about the same at this point, that's significant for Gemini. So anyway, Google is not to be left behind. They're still there. And have you much experience with using Gemini?

> **Leo:** Yeah, yeah. It's pretty good. I mean, they've been kind of behind; right?

Steve: They have been, yeah, we have not been talking about Gemini that much.

> **Leo:** Yeah. So, yeah, I mean, I've been using it. It ain't bad. It's not my go-to by any means.

Steve: But it's not your go-to, yeah.

> **Leo:** I think they get a lot of users by default because it's what powers their search.

Steve: They're Google, yes.

> **Leo:** They're Google. And it's what powers Google Docs and, you know, the Google Workspace.

Steve: And it's a button right there in the Google browser bar.

> **Leo:** They're about to get a massive boost in users tomorrow because Apple's using it for their Siri AI.

Steve: Well, thank god. I mean, they're using at least a good AI.

Leo: It's better than Apple's. It's better than Siri.

Steve: Oh, anything is better than that, yeah.

Leo: I've been using the beta, and just the dictation alone is light years better. The dictation was so stupid before. It's now fairly decent. But Google is not cutting-edge, or at least they have - they don't have the reputation of being cutting edge. And while this Gemini 3.8 Flash is - it's pretty good.

Steve: It's better than 3.7 three weeks ago.

Leo: Yeah. It's good at prose. I haven't tried it with coding. It's pretty competitive.

Steve: So not a strong enterprise buy. You know, enterprises are OpenAI and Anthropic.

Leo: It might be because it's Google. Yeah. I mean, yeah, it depends. I don't know why Google has stumbled so much in this.

Steve: Interesting.

Leo: Yeah.

Steve: Before we wrap up the AI-related news, I want to share a summary that Risky Business News produced from a much longer report by Bitdefender. Since I always try to go to the source I tracked down Bitdefender's write-up. But it contained so much superfluous information that I returned just for Risky Business's much shorter summary which still offers the information that we care about.

So here's what they said. They said: "A new report" - meaning the Bitdefender report - "describes how a Chinese cyberespionage outfit is using AI to beef up its malware arsenal. If this is a sign of things to come, clustering threat actor behavior together for attribution purposes is about to get a lot more difficult." By that they mean all the security firms are able to do is guess at who did what based on the behavior that they see. So they'll, like, notice, oh, you know, a lot of this code sure does look a lot like a lot of that code. So these are probably related to each other somehow. I mean, you were literally left to kind of piece the forensics together that way.

And so what Risky Business starts by noting is that what this Chinese espionage outfit is using its AI for is going to make this more difficult. They said: "The Bitdefender report, released last week, describes seven remote access tool (RAT) families. All seven were created by a single cyberespionage actor Bitdefender called SilkParasite, and five - I know - and five were previously undocumented." Meaning never been seen before. "The report authors have medium confidence that SilkParasite is a 'China-Nexus actor' targeting governments across Central Asia including Uzbekistan, Turkmenistan and Kazakhstan.

"Back in November," they say, "we wrote about what looked like an experiment to see how AI-assisted hacking could support China's Ministry of State Security. The approach those threat actors took at the time was to build an attack framework and let Claude do the hacking. It was error-prone and noisy, and sometimes successful. SilkParasite, by contrast, is not using AI for Yolo hacking. It is using it to support cyberespionage programs where important goals include operating stealthily and not getting caught.

"According to Bitdefender, SilkParasite 'develops, tests, debugs, and iterates its own tooling, maintains a structured build and deployment workflow, and regularly rotates infrastructure, encryption material, payload names, and persistence artifacts between deployments.'" They said: "Rotating its infrastructure and indicators of compromise between deployments makes it harder to detect and link its activities together." Essentially, these guys are using AI to become far more stealthy so that their individual instances of attack look like separate non-commonly attributable attackers.

They wrote: "SilkParasite also uses a variety of programming languages and command-and-control protocols. Its seven different RATs are written in different languages including .NET, C++, Go, or JavaScript. Command-and-control is accomplished by abusing Google Drive and Internet communication protocols including HTML, HTTP, TCP, DNS, and TCP.

"Its malware also typically uses a modular plug-in architecture where additional functionality is only deployed when it's needed. This means initial implants are relatively small, and plugins are used to provide capabilities for clipboard monitoring, keylogging, file management, or interactive shell access. This limits the exposure of the entire toolset during any single deployment. It also allows SilkParasite to update or replace individual components without having to change the entire implant. And it minimizes the amount it writes to disk, typically only the files required to get its malware up and running. For victim organizations, these measures make forensic analysis more complicated. Complete remediation is also more difficult once a particular implant is detected."

Okay. And remember that one of the key features that we're seeing now in cyberdefense is the so-called IOCs, the Indicators of Compromise. But if every use has different indicators, then nobody who's keeping track of all the latest seen, you know, previously seen indicators of compromise will have their alarms tripped when what looks like a brand new one-off attacker shows up. So these guys are very cleverly using AI to basically maintain an amorphous presence and to be a chameleon attacker.

They said: "To us, all the behaviors above are the hallmarks of a professional cyberespionage outfit. Of course, doing all of this in a disciplined way is a lot of work, and Bitdefender has evidence SilkParasite is using AI to help it deliver this complex engineering." Yeah, I mean, this is a different scale of engineering than we've seen from bad guys before.

"SilkParasite's malware contains indicators of AI-assisted development, such as left-over test functions and placeholder encryption keys. Intriguingly, implants that Bitdefender dubbed GoginRAT and NomadRAT share a high-level architecture even though they are written in Go and C++ respectively, and use different command-and-control protocols and code structure. Therefore, Bitdefender suspects that the same high-level specification document" - meaning, you know, prompt - "was independently implemented twice with AI assistance. Bitdefender concedes that this structural similarity is not conclusive evidence, but notes that it is the 'kind of thing an AI-assisted workflow makes very easy.'

"This is the first example we've seen where the evidence tells a compelling story of a competent cyberespionage actor incorporating AI into its work practices. SilkParasite is taking the same disciplined approach to malware development and doing more of it. It's creating more malware families to build redundancy, make attribution and discovery harder and reduce the risk of compromise from any single exposure.

"Bitdefender has done a good job describing SilkParasite's malware families and has published Indicators of Compromise. That kind of exposure would once have set the group back significantly." The group meaning SilkParasite. "Now that they've figured out how to use AI to speed up their development work, they'll be back better than ever relatively quickly. Then the discovery, attribution, and publication merry-go-round can start all over again."

So for me, what's reported here presents a sane, entirely defensible and believable example of the way AI will be used to conduct offensive cyber operations. Apparently it's already happening. Certainly it will be in the future. So much of what we're hearing about the coming AI-driven cyberpocalypse, when rogue AI agents freely roam the Internet wreaking havoc wherever they choose, none of that makes any sense to me. What I expect to see is pretty much what we've been seeing, only more broadly and deeper. Governments will use AI to infiltrate their espionage targets,

and criminal gangs will use it to infiltrate commercial enterprises, then exfiltrate their valuable proprietary data, then attempt to extort and blackmail them using the data they obtained.

I just don't see any coming cyberpocalypse. Yes, everyone needs to be more wary, and the probable targets of these operations, governments and commercial enterprises storing data they must protect, need to shore up their cyber defenses. That's absolutely true. You know, pay more than passing lip service to their CISOs. You know, make everything as secure as you possibly can because we are going to see a much increased obvious increase in the use of AI.

Okay. This is the last of the two things I wanted to talk about, and I'm excited about this because I think this is really interesting. This is, I think, an interesting thought piece posted by one of, as I said, our favorite cryptographers, Johns Hopkins Professor Matthew Green.

A couple of weeks ago Matthew proposed something interesting that I think is worth pondering. His blog's title was "Everything is about to 'go dark.'" His piece explores the possible consequences of AI being used as it certainly is and is going to continue to be to make many systems far more secure. Nobody could argue that if Microsoft has just patched nearly a thousand bugs, that Windows is better off today, after today's Patch Tuesday, than it was yesterday, on Labor Day. So Matthew suggests there may be unforeseen consequences of that. So here's what he wrote.

He said: "I'm coming down from spending a few days at USENIX Security, right here in my hometown of Baltimore. This means that my days have been taken up with two kinds of conversation: first, explaining to my colleagues why Baltimore is not actually like 'The Wire,' which of course is HBO's famous series which is fantastic."

Leo: I hope not. Yeah.

Steve: He said: "And second, trying not to talk about AI." He said: "I'm going to break that second rule now." He said: "I have" - this is Matthew Green saying "I have many worries about what AI means for our field, for various definitions of 'field.' But in this post I want to focus on just one thing I've started worrying about, and it's a perverse thing. Specifically, I'm concerned that AI is going to make software much too secure.

"While that doesn't sound so bad on the surface, there's a consequence to this," he says. He said: "I mean something very specific: I'm concerned that U.S. intelligence and law enforcement agencies are about to go dark, meaning that they're going to suddenly lose a huge portion of their capability. And that this isn't going to be simply a problem for those agencies, but also for those of us who value computer security and privacy in general."

Okay. So what does "going dark" mean in the era of law enforcement hacking? He said: "To explain how we got here, we need to talk about recent history. This actually gives me a real excuse to reference The Wire, just because it's a perfect snapshot of what electronic surveillance looked like way back in 2002. If you've seen the first season, you'll recall that it's about cops wiretapping drug dealers who use payphones and burners."

Leo: Burner phones, yeah.

Steve: Yup. "The mobile phones in the show are relatively new technology for the time, but from a technological perspective nothing in this scenario would have shocked a cop who jumped forward from, say, 1989. The change began in the late 2000s, thanks to the rise of smartphones and texting. Because smartphones can actually store data as well as conveying it, the contents of those phones quickly became a useful new source of law-enforcement capability. Or they were until 2010, when Apple began encrypting iPhone storage using a key derived from the user's passcode."

And he says: "Android phones followed shortly thereafter. The next year, Apple deployed end-to-end encryption in iPhone text messages. By 2014, a tiny texting startup named WhatsApp had gathered 600 million users worldwide. By 2016 those users, now nearly a billion strong, were all using default

end-to-end encrypted messaging and calls. These two trends the move from calls to texts, and texts to encrypted data happened very rapidly.

"The FBI and law enforcement agencies were not insensitive to what was happening. In 2014, Director Comey announced an initiative called Going Dark, which would launch a 'national conversation' about what providers could do or be compelled to do to make these new communications media legible to law enforcement and counterintelligence.

"In 2016, the agency quit talking and took their theory to court. When a terrorist attack left the FBI holding a shooter's locked iPhone, the agency ordered Apple to give them access. The company refused. What broke the stalemate and, to some extent, ended the Going Dark conversation was something neither the FBI nor Apple expected. An outside company announced that there was no need for Apple's assistance; they could simply hack the phone.

"The Apple v. FBI case has turned out to be a microcosm of the whole Going Dark debate. For the next decade, law enforcement and intelligence agencies continued to ask for 'exceptional access' backdoors, but the urgency was gone. Agencies and manufacturers both knew that law enforcement could and would purchase targeted hacking tools like GrayKey (for phone unlocking), or even remote exploitation tools like NSO Group's Pegasus, if they needed them badly enough. Vendors like Apple and Google continued to play a vigorous defense, closing vulnerabilities as soon as they learned about them. But offensive vulnerability hunters consistently managed to keep the edge. But now, today, there's a very good chance that all this is about to be history because the era of AI bug hunting is here.

"This April, just four months ago, Anthropic announced a new model called Mythos that happened to be unusually skilled at software vulnerability discovery. The U.S. government temporarily blocked its export, restricting access to U.S. agencies and trusted vendors. While the ban was dramatic and made for good PR, it turned out to be mostly pointless. OpenAI, along with Chinese open-weight model labs like Z.ai and Moonshot, have since demonstrated that vulnerability finding is not something that a single lab is likely to hold a monopoly on. The growing list of serious vulnerabilities these models have found is getting scarier and more impressive by the day.

"At first glance, this might seem like good news for the offensive team, and for hackers in general," he says. "But I doubt that's how this will play out in the long term. Defenders are now in the process of patching every bug they can find, often decades' worth of bugs, and the backlog feels huge. But they're making progress. Entire CI toolchains are being rebuilt to incorporate AI-based vulnerability scanning before software ever reaches the point where a human will touch it. While I doubt this means that every bug will be found, in the real world, it does feel likely that we're going to hit some sort of a ceiling on the number of useful bugs, and we'll probably hit it soon."

So in this regard Matthew and I are in complete agreement. We're going to see this bug discovery and patching rate eventually drop, and drop near to zero. He says: "Thus, over the next two years, major pieces of software are likely to run out of remotely-exploitable bugs."

He says: "Obviously, I think this is great. But for law enforcement and offensive intelligence agencies, it's going to be a nightmare. For the first time since 2010, law enforcement might experience what it looks like to really 'go dark,' across a huge category of advanced, well-maintained devices and pieces of software. So how is this a problem?"

"The debate over 'exceptional access' mechanisms never really went away. In some places, like the UK, it even metastasized into something worse. Here in the U.S. it mostly went into hibernation. Some of the slowdown can legitimately be attributed to expert pushback - academics and industry engineers pointing out the risk that backdoors might be abused by the very adversaries that agencies are supposed to be protecting us against. But I fear," he writes, "that this was less of a principled pause, and more of a market that was just pricing supply. The destruction of the low-hanging vulnerability fruit will make law enforcement and intelligence agencies' needs much more acute. The demand for constructed, intentional backdoors will re-start in earnest.

"The result will be enormous pressure on industry to re-architect their systems to make their systems amenable to exceptional access. In some cases, governments will ask for these capabilities in the expectation that they'll be useful for spying on other governments a strategy that might have been undetectable in the pre-AI era, but that probably will be less productive now. The results are

unpredictable. One result might be that non-U.S. governments entirely remove their dependence on U.S. software.

"The worst part about this dynamic is that these potential new backdoors will probably only affect the countries that demand them, meaning that they will be primarily useful for allowing the U.S. to weaken its own systems. This will in turn allow foreign adversaries to find new ways to attack our communications. This deliberate self-sabotage will happen just at a moment when we're finally getting a handle on securing our own infrastructure.

"So what do we do about it?" He says: "I honestly have no idea. This is not a call to action for experts to rally behind a sophisticated plan. Like so many things about the AI revolution, it's just occurring to me that we're on a long greasy slide to a place that will look different than where we are today. Just realizing this doesn't mean I have any strategy in mind to avoid it. In this case, we're just going to have to hope that this time we make the right choices, for no other reason than they're right."

So I wanted to share this because I think it's a brilliant forward-looking take on our near-term future. Over here in the U.S. we've been watching the United Kingdom and the European Union wrestling with this issue over and over. And we've comfortably watched it refuse to die. While comfortable from a distance, watched it refuse to die. It just will not die. But this very issue may soon be visiting our shores here in the states. We've watched our United States enact laws that have forced adult content websites to black-out access across entire states because there's currently no practical way to guarantee the age of anyone visiting.

And at this very moment Utah's Senate Bill 73, which was signed into law on March 19th and which sailed through Utah's state government passing 22-to-2 in their Senate and 66-to-1 in their House, meaning a stunning majority, was written to take effect last Thursday, September 3rd, but was just extended by 14 days, two weeks, to September 17th, Thursday after next. THAT onerous and entirely unworkable Bill deems a person who is physically located in Utah to be a Utah user, regardless of the IP address they present to any age-restricted Internet service.

Its shorthand is the anti-VPN legislation because it's clearly meant to curtail the use of VPNs and other proxies as a means of geo-relocating. No one has any idea what's going to happen there, but it's going to be interesting to watch because, you know, basically they are trying to prevent adult age restricted Internet service websites from allowing connections from VPNs. But not all VPNs declare themselves as such.

So my point is, there is a clear tension growing between the legislatively protected privacy rights of citizens, and in some cases we have constitutionally protected privacy rights of citizens in many major democracies and the perceived needs of their own governments to conditionally violate those rights. In an analog and pre-encrypted world, law enforcement and intelligence services were able to sneak around to get what they believed that they needed. And even after nominal encryption was emplaced, until the advent of AI, the large supply of latent bugs allowed these same agencies to ignore privacy whenever they felt it was necessary to do so. Matthew's point is, that fragile status quo is soon to end. What will replace it?

Leo: Good question. He's a smart guy. And he's very AI-aware, too. He's actually a pretty pro AI guy. So...

Steve: Yeah. And he's also one of the guys that, you know, that the legislators pull into for Senate hearings to find out what he thinks.

Leo: Yeah.

Steve: And, I mean, so here we're saying look how much more secure Windows is today than it was yesterday with a thousand new problems fixed.

Leo: Ooh, boy.

Steve: Yeah. And we know that Android and iOS are going to be on, and macOS are going to be on - they're all going to be on the same curve. They're all going to be getting the benefit of this, and we could argue that AI is going to help future errors, as Matt said, not get into production code. So we're going to fix this. And, you know, Apple has struggled to keep their phone from being hackable. It's been a struggle. They're probably going to win. And then what?

Leo: Well, I mean, they've been complaining about going dark, as you pointed out, since James Comey. And yeah, digital technologies have given them more ways of seeing into our lives than ever before.

Steve: Exactly. And people like the NSO Group with Pegasus have been able to keep compromising people's phones.

Leo: Yeah.

Steve: What happens when that changes?

Leo: When they can't. Well, then we're back to the way it used to be when law enforcement didn't know every darn thing that was going on inside your house, inside your mind.

Steve: Except that we were talking over analog phone lines and wondering if that little click and static sound was somebody listening.

Leo: There's no question law enforcement is always going to want more, even if they weren't going dark. They're always going to be pushing for this. Right? This is just - they have been, and they will continue to push for this.

Steve: Well, the problem is our legislators are now going to, you know, they're going to come under pressure to legislate this.

Leo: Right. I see this as part of a much larger uncertainty in general. We are entering into - it's really, I would say, fairly chaotic.

Steve: We are in a time of phenomenal change. I agree, Leo.

Leo: One thing we know about chaotic systems is they're very hard to predict. It's just not - they're not deterministic.

Steve: Yup.

Leo: And I think this is chaos.

Steve: Yes.

Leo: We don't know what AI is going to do. If the scale ranges from it's just more computer programming to it's an alien intelligence in our midst, and I don't know where it's going to land on that scale, and I don't know what the disruption's going to be. Will people lose their jobs? I don't even know if that's clear. It's all very through a glass darkly. So I don't - I just don't know if we can make any sensible plans I guess is what I'm saying. We should be - there's a storm coming. And I don't know what you do to prepare for it except maybe get more rice. I don't know.

Steve: Buckle up.

Leo: Buckle up. It's going to be a bumpy night. Pretty weird.

Steve: Okay. Last break, and then I am very excited to share this lifelong engineer's observation about the danger of becoming overly reliant on automation.

Leo: Yes.

Steve: And we have a new automation capability in town. AI.

Leo: AI-i-i, as I have been saying. Now on we go with Mr. G.

Steve: Okay. A guest-submitted article recently appeared in the IEEE Spectrum publication. It captured my imagination, and it feels very important to me. And I believe it's going to resonate deeply with many of this podcast's listeners who've been around the block a few times. For those who don't know, IEEE is the abbreviation for the Institute of Electrical and Electronics Engineers. It was founded as the AIEE, the American Institute of Electrical Engineers, believe it or not, in 1884, an astonishing 142 years ago. And then it was later renamed to IEEE after its 1912 merger with the Institute of Radio Engineers, the IRE. And, you know, to put the Institute's age into perspective, it was at the start of the year of its founding that a penniless genius by the name of Nikola Tesla arrived in New York to work for an already famous industrialist by the name of Thomas Alva Edison. But anyway, I digress.

Today, the IEEE describes itself as the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity. And the article I encountered which so galvanized me carried the headline: "AI Efficiency Could Cost Us the Next Generation of Experts." And the article's teaser read: "Lessons from aviation and nuclear power show how to preserve human skills." And as I said, I believe what's said here is extremely important. So see what you think.

Its author wrote: "A little over a decade ago, I led the controls design for a first-of-its-kind full digital control system for a U.S. nuclear plant. It was, on paper, a beautiful machine, engineered to run itself the way a modern airliner does, with operators watching over a system that rarely needed them. And we made a decision that, to an efficiency-minded observer, looked backward: We deliberately left manual steps inside sequences the system could execute on its own.

"We were solving a specific problem. An operator who only ever supervises automation slowly stops being an operator. The hands go cold. The mental model of what the plant is actually doing gets fuzzy. Then comes the day the automation hands control back. It's always the worst day because automation only quits when it's confused or in trouble. But by then, you have a person in the chair who has not truly operated the thing in years. The manual steps we included in its design were there to keep the human current. It was inefficient by design, on purpose.

"The plant, as it happened, was never built. It was shelved amid the politics and economics that surround nuclear power in this country, for reasons that had nothing to do with the engineering. But the design instinct outlived the project, and I've become," he wrote, "I've come to believe it's the

most useful idea I can offer to the argument now consuming every boardroom: What happens to human expertise when AI does the work that used to build it?

"The data has become hard to wave away. A Harvard University working paper covering some 65 million workers at more than 280,000 U.S. firms found that after companies adopted generative AI, junior employment fell roughly 9% within six quarters, relative to non-adopters, while senior employment kept right on growing. A Stanford analysis of ADP payroll records points the same way: The youngest workers in the most AI-exposed occupations lost ground after late 2022 while their more-experienced colleagues held theirs. The Stanford researchers found that the losses concentrate where AI automates the work; where it merely augments, junior employment holds steady or rises.

"The causal story is still contested, and honesty requires saying so. Researchers at the New York Fed attribute much of the rise in young-graduate unemployment not to AI, but to remote work, arguing that firms are reluctant to hire inexperienced people whom they cannot train and mentor at a distance. But notice what the explanations share. Whether a model is absorbing the formative work or distance is severing the mentorship around it, both describe the same broken mechanism: the apprenticeship channel through which expertise passes from senior to junior. Either way, 'entry-level' has quietly come to mean 'three years of experience required.'"

He says: "Strip away the noise and you're left with one deceptively simple problem: You cannot become a senior engineer without first being a junior one. Expertise is not downloaded. It is earned through failed builds, dead-end debugging sessions, and the 'why on earth did that work?' moments that a capable AI will now happily spare the newcomer. Spare them enough of those, and you produce a cohort that can supervise a model on paper but never develop the gut sense to know when the model is confidently, catastrophically wrong.

"Most of the commentary stops at the diagnosis, or reaches for policy solutions that treat the loss of junior jobs as an economic problem. Yet it's also an engineering problem, and safety-critical fields have already spent decades learning how to solve it."

He said: "My own career began at the sharp end of automation. My first job out of school was verifying and validating the software in the digital jet-engine controller that decides, faster than any pilot could, how a fighter plane's engine responds. Even then, in the late 1980s, the central tension was visible: The machine outperforms the human in routine cases, but the human is all that stands between the aircraft and disaster in the cases the machine did not anticipate. This tension is known as the automation paradox, in which increasingly capable automation gives human operators less practice, while leaving them with only the most difficult situations.

"Aviation learned, repeatedly and expensively, what happens when human skills atrophy inside that gap. The canonical example is Air France Flight 447, which fell into the Atlantic in 2009. The proximate cause was mundane. Iced-over airspeed sensors fed the autopilot bad data, and it did what it is designed to do: It disconnected and handed control of the airplane back to the crew. What followed was not a hardware failure. It was a competence failure. A recoverable situation became an unrecoverable one because the pilots, conditioned by thousands of hours of watching the automation fly, could not read a high-altitude aerodynamic stall and hand-fly their way out of it. The airplane was working. The training the automation had quietly eroded was not.

"The industry's response is instructive, and it's the same move we made in that nuclear control room. It did not rip out the autopilot, but built deliberate manual practice back in. In 2017 the FAA issued Safety Alert for Operators 17007, 'Manual Flight Operations Proficiency,' declaring that 'manual flight is the foundation upon which other technical flying skills are built.' The alert formally recognized skill decay as a hazard in its own right. Some airlines amended their procedures to encourage hand-flying both the initial climb and initial descent in benign conditions, knowingly trading a sliver of fuel efficiency to keep the crew's raw flying skills alive. That trade is the whole point. A perfectly optimized system that produces incompetent operators is not optimized at all. It has simply moved its failure mode somewhere the spreadsheet cannot see.

"Put the aviation lesson and the nuclear instinct side by side and they point to one design pattern we now need in AI-augmented work: the deliberate 'manual gate.' A manual gate is a point in a workflow where a human takes the controls, not because it is the fastest way to get the task done, and not as a safety interlock, but specifically to exercise and preserve a skill that would otherwise decay. The distinguishing feature is that it is chosen. You decide, as a matter of design, which

competencies your organization must keep alive in human beings because those are the ones you will need on the bad day, when you engineer the friction required to keep them warm.

"Picture how this might work on a software team that leans on AI for most of its code. The team places a manual gate around the skill it can least afford to lose: debugging. When a defect surfaces in a critical module, the assigned engineer - deliberately often a junior one - must first reproduce the failure, trace it to root cause, and write an automated test that captures the bug, all with the AI assistant switched off. Only after the engineer commits to a diagnosis does the model come back on, to propose the fix, generate alternatives, and sweep the code base for similar bugs. The engineer then compares their diagnosis against the model's. When the two disagree, that's the design working, surfacing the disagreement before the bad day instead of during it. It's also the design teaching and training the junior engineer.

"This approach reframes the junior engineer entirely. The instinct today is to let AI do the entry-level work because it is faster, cheaper, and capable. But some of that work is not overhead to be eliminated. It is the training apparatus of your future senior staff, and you should protect it the way you'd protect any other piece of critical infrastructure. It may not be efficient this quarter, but dismantling it quietly mortgages your capability a decade away.

"None of this is free, and pretending otherwise would insult the people who have to sign the budgets. A deliberate manual gate is, by construction, less efficient in the near term than full automation. Keeping juniors doing formative work and running the manual sequences costs something now to protect something later.

"That's a hard sell in a market that judges most leaders on quarterly results. A hired executive who carries [he has in air quotes] 'unnecessary' humans that AI could replace will hear about it from the board long before the payoff arrives. The math only works for someone insulated from that pressure: a founder with control, a private company, an institution with a genuinely long horizon, or a regulator willing to require workers to demonstrate their skills regularly, as pilots must. This all means the organizations most likely to preserve their own expertise are the ones structurally able to spend short-term margin on long-term capability; everyone else will need a push from the outside.

"So here is the argument, in one line: Deliberate inefficiency is not waste. In safety-critical engineering we have always known it as insurance, and we buy it on purpose. As AI takes over the work where expertise is forged, the smart move is not to resist the automation. It is to keep our hands on the controls by design so that, when the automation fails, as it always eventually does, there is still someone in the chair who knows how to fly."

So I think that's a fantastic piece of well-reasoned engineering, and I would recommend it without hesitation to anyone's boss who may currently be enraptured by AI's truly deliverable ability to eliminate all of those pesky junior engineering jobs.

You know, we already had stunningly good AI with GPT-5.6, Fable, and Mythos. And though it's going to take some time for the world to fully assess what OpenAI has just given us with GPT-6 Astra, the early take is that it currently outperforms some of the other frontier models from OpenAI, Anthropic, and Google. Whether or not and to what degree that's true, and even the fact that this is nothing more than a snapshot in time, that's important because of course this wasn't true a week ago, but may be true today. And we know that we are just at the beginning of all this. We can feel how rapidly this field is changing. And so that's my point. If we've learned anything, it's that no one will ever have a long-term defensible position in AI. Sure, you know, they're going to be bouncing back and forth. And Leo, you already knew that Anthropic was, what, a week away or so from releasing their next model in this never-ending series.

> **Leo:** Supposedly. I mean, it's rumored, anyway, that they are ready with Fable 5.1, yeah.

Steve: So anyway, I think without the deliberate creation of an environment that is designed to nurture and mature junior developers, organizations will be left completely dependent upon automation for mission-critical functions.

Leo: Yeah.

Steve: I mean, I know that coders who come out of university with no experience, they're just not useful. They actually can't do anything. And you could argue that they're going to be a lot less useful. All they'll know how to do is drive AI. No coding. I don't even know if programming will be taught in the future. It'll be prompt engineering.

Leo: Right. And you think that we need to know how to code.

Steve: We probably need to know how to read what AI does when it writes a bug. I don't know yet whether we're going to be able to say to AI, fix the bug which you created.

Leo: Right.

Steve: Maybe we will. I mean, so I guess the question is, okay, so certainly there will, well, AI is also coding itself now; isn't it.

Leo: Yeah.

Steve: So maybe it will be a completely lost art except here in Irvine.

Leo: Yeah. Yeah, right. There'll be a few people. I mean, I think there'll always be a few people who code by hand, just like there are a few people who make their own furniture.

Steve: And play the piano.

Leo: Yeah.

Steve: Instead of, you know, turning on Apple Music.

Leo: Yeah. I just, I don't - code is a funny thing because code is getting the computer to do something by translating your English language thoughts into machine code. Of course [crosstalk] plus the machine code. Your intent.

Steve: It's an unvoiced - your intent. Because it might be unvoiced.

Leo: Right. I mean, all - and as we've moved higher up, you still do low-level coding, but most people are doing high-level coding, which isn't really that closely related to what's going on in the computer. I guess there are, there is a case to be made for people who need to be able to look at the traces and see what's actually happening.

Steve: Well, for a long time, compilers had bugs. And which meant you would properly express yourself in the high-level language, and you still wouldn't get the program doing what you told it to do.

Leo: Right. This to me is another case of we don't know what the hell's going to happen because...

Steve: Well, there are people talking about AI slop code, and that we're now talking, and we're going to be introducing a whole new generation of bugs. I don't know one way or the other.

Leo: I think that's old farts talking.

Steve: Could well be.

Leo: Yeah. I think that that's people who are very tied to the old way of doing things. Slop is a pejorative that I don't think is completely fair. It doesn't code the same way you do, that's true. But the thing about computer programs is, if they're correct, they work. The problem is these little edge cases where it works most of the time, doesn't work all the time. But I think a computer is ultimately, at least I don't see any reason why a computer wouldn't ultimately be the best at figuring out what's going on. Right? It's better at it than we are.

Steve: I did, I was arguably one of the early people to say that AI was going to be extremely good at code.

Leo: This is what it does.

Steve: Exactly.

Leo: So really what the AI is doing that's hard for the AI is is it good at us. And I think most of the time when the AI fails us, it's because it's not as good at us, or we're not as good at it. It's a communication problem between us and the AI, a misunderstanding or miscommunication or misdirection. I mean, I literally haven't looked at code in a long time. And I think most - many of the people like Darren, who's a very proficient coder in our Discord, I looked up to him because, you know, he's the guy who would finish all the Advent of Code problems lickety-split. So he's a very, very accomplished programmer, done it for years, his whole life. He says: "I haven't coded anything myself for like a year."

Steve: Wow.

Leo: And I think many of the most proficient coders I know, people like David Heinemeier Hansson, don't look at the code. Now, his operating system Omarchy, which is a version of Linux that's entirely AI coded, is kind of wonky. But that doesn't - I'm not sure I blame the AI. I mean, many of the coders who are using AI at the point have stopped looking at the code. So I don't know what that means. I don't know. I think Darren's saying you can absolutely build great quality software without knowing anything about writing code right now. And I think that's only going to get better.

But there needs - but it's absolutely true there needs to be entry level positions. You know, I think the analogy is to the business I'm in, broadcasting. You're terrible when you start. You really are. You're not good at it. And you need somewhere to start. You need basically - it took me a year to find a radio station that was so tacky and small that it would let me learn how to be a broadcaster there. And they put me on in the middle of the night, midnight to 6:00 a.m. on Sunday morning, because that was where I could do the least harm. But it gave me a chance to learn.

Steve: And we know it's called apprenticeship.

Leo: Right.

Steve: The idea is you apprentice with a master, and you learn his art.

Leo: Right, right. But then I look at my son, who really has no broadcast training, who just did it all in public on TikTok and has done a little bit - has done pretty well for himself. A lot better than [crosstalk]. And he did it, by the way, in a hundredth of the time his dad did. I mean, this is literally, he's gone from nothing to total success in whatever this business is, sort of like broadcasting, in a couple years. Took me 20 years - 10 years to get halfway decent. So it's a different time. And I think a lot of what you're seeing in a lot of this complaining and caviling about AI is our older generation going, well, that's not how you do it. In my day [crosstalk].

Steve: How do you know what's in the registers? You've got to know what's in the registers.

Leo: Yeah. Do you care what's in your registries? You don't. You do. And you know intimately. But I don't know if you need to. And certainly somebody who's programming in C++ probably doesn't.

Steve: No. You don't have any access unless you explicitly tell it you want to.

Leo: Yeah, you don't really care.

Steve: A bit of MASM.

Leo: You don't care. So I don't - I just don't know. We're in unknown, uncharted territory.

Steve: We're riding a beast.

Leo: It's a tsunami headed for us.

Steve: Yeah.

Leo: And there are a number of people who are going, "Run," and a number of people going, "Oh, look, the water's going out. I see seashells." And I think I might be in that latter group, and the wave's going to hit me. Steve, always, always provocative, intriguing, entertaining, and informative. It's a great show, and I really appreciate your putting all this work in you do every week.

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/