SECURITY NOW!

Transcript of Episode #1027

Artificial Intelligence

Description: What the status of Encrypted Client Hello (ECH)? What radio technology would be best for remote inverter shutdown? Some DNS providers already block newly listed domains. Knowing when not to click a link can take true understanding. Why can losing a small portion of a power grid bring the rest down? Where are we in the "AI Hype Cycle" and is this the first? Speaking of hype: An AI system resorted to blackmail? Why are we so quick to imbue AI with awareness? ChatGPT's latest o3 model ignored the order to shutdown. Copilot may not be making Windows core code any better. Venice.AI is an unfiltered and unrestrained LLM.

High quality (64 kbps) mp3 audio file URL: http://media.GRC.com/sn/SN-1027.mp3
Quarter size (16 kbps) mp3 audio file URL: http://media.GRC.com/sn/sn-1027-lq.mp3

SHOW TEASE: It's time for Security Now!. Steve Gibson is here with some shocking tales from the frontier of AI. He'll also talk about the status of Encrypted Client Hello. It seems like a good idea. Maybe it's not. And what radio technology would be best if I wanted remote shutdowns of solar power inverters? All of that and more coming up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1027, recorded Tuesday, May 27th, 2025: Artificial Intelligence.

It's time for Security Now!, the show where we cover your security, your privacy, how computers work, and a whole lot more with this cat right here. The cat's been drinking some milk.

Steve Gibson: Got it in his moustache.

Leo: It is Steve Gibson. No, don't do - he's got a lovely little froth catcher device on his lips.

Steve: I do.

Leo: Yes.

Steve: And it needs to be independently cleansed.

Leo: I don't know want to know how.

Steve: No, you don't want to hear about that.

Leo: So today the show's all about AI? No.

Steve: No. Well, no. As things came together, there were a number of stories that were about privacy and security, and I ended up with maybe four or five different things. And I thought, okay, I think we should just - I'm just going to title this "Artificial Intelligence." For a while, as I said to you before the podcast, I had Artificial Artificial Intelligence, and I thought, no, that's a little too cutesy-poo, so we don't need to do that.

Leo: You know, you could use AI to design the title. Just ask AI for an artificially intelligent title.

Steve: I could probably do that.

Leo: Shall I look up some candidates for you?

Steve: And I'll tell you, I don't know if you've spent any time with the o3 model; but whoa, is it astonishing.

Leo: I think there's some really good ones out there, yeah.

Steve: I don't know what - oh, wow. Anyway, we've got a bunch of stuff to talk about. One of the reasons that, well, there was no other overriding central event, although I actually already have some things to talk about cued up for next week that I will talk about at the end of this podcast. But I wanted to talk about the status of Encrypted Client Hello, ECH, which we did a podcast on entirely, one of our deep dives, about 18 months ago, toward the end of 2023. I ran across a security group in the business that noticed some recent changes and thought, what?, and then did some of their own deep diving that I want to share.

There was a bunch of - the feedback from our listeners has become just golden, which is why some of our podcasts have had a lot more of that. So they serve as talking points for answering questions like what radio technology would be best for remote inverter shutdown. Turns out that some DNS providers already do block newly listed domains, which is something we were talking about last week. It became clear that knowing when not to click a link can actually take some true understanding. Many people need to keep asking because, you know, they're not the people who listen to this podcast, essentially.

Also, why can losing a small portion of a large power grid bring way more of the grid down than one would think? Where are we in the AI hype cycle, and is this the first? Maybe it's the last. We'll see. Speaking of hype, an AI system resorted to blackmail? What? Also, why are we so quick to imbue AI with awareness? It turns out that goes back a long way, way further than most people might imagine. Also ChatGPT's latest 3o model ignored its orders to shut down, which raises some interesting new eyebrows. We're going to run out of eyebrows, Leo. And also it turns out that Copilot, which

Microsoft is now using with great abandon, apparently - remember we recently heard that, what was it, a third of the new code is being written? Turns out that it may not be making Windows code any better.

Leo: Yeah, what a shock, yeah.

Steve: We've got some details from GitHub which are a little concerning. And I stumbled upon an unfiltered and unrestrained LLM that I thought our listeners would probably have some fun playing with.

Leo: Yeah.

Steve: And of course we've got a Picture of the Week that has already generated a week's worth of feedback from the people who received all the show notes yesterday around this time. It's surprising to me how many people have had the events surrounding this picture occur in their lives.

Leo: Oh, wow.

Steve: Yeah.

Leo: Now, I haven't seen it because I put myself in a soundproof booth the minute your email comes in. So I am virgin ears and eyes.

Steve: You and Lorrie's son impress me with your speed of acquisition. I showed this to her son a couple days ago, and you would measure in milliseconds the amount of time it took him to process the photo and burst out in a belly laugh. And I thought...

Leo: He's a smart kid. Really smart kid. He's probably not a kid, but he's smart, anyway.

Steve: Yeah, definitely.

Leo: This is - I did feed your show description into the AI, and these are the titles it came up with: "AI Hype Hysteria" and "Hard Truths."

Steve: Ooh, I'm going to do that from now on, Leo.

Leo: AI: "Hype, Hacks, and Headaches." "AI Security Hype vs. Reality." "When AI Crosses the Security Line." This is my favorite. "AI: Friend, Foe, or Flaw?"

Steve: Oh, Leo.

Leo: There's some good ones in here.

Steve: Which AI? Oh, Perplexity, huh?

Leo: Yeah. I love Perplexity. I'm a big Perplexity fan.

Steve: That really is really good.

Leo: Yeah. Well, we will get to the show. Whatever it's called, it's going to be good.

Steve: Yeah, now I'm annoyed I called it "Artificial Intelligence."

Leo: You could have used it. That's the new "I could have had a V8." Could have used AI.

Steve: That's right, I could have used AI.

Leo: Man. All right, Steve. I am going to go to the show notes.

Steve: So I gave this picture the caption "Who could possibly believe this was a good idea?"

Leo: Who could possibly believe? And I am going to - you could start the timer - scroll up and see a water heater floating in the air because of course floor space is a premium.

Steve: Premium, yeah.

Leo: What's below it? Yeahhhhh. It even says "Danger, High Voltage, Keep Out." And then, really to add insult to injury, there's some mop buckets there, too.

Steve: Yeah. This picture just keeps on giving. For those who are not privy to this photo, basically the picture is in two halves, an upper and a lower. Unfortunately, the upper is the upper as opposed to the lower. On the bottom of the photo, where a lot of the potential action will be happening, is a lineup of large high-tension, high-power, high-voltage breakers with those levers on the side for pulling the master switch. Looks like we've got some big power, like, trans - substation kind of transformer things.

Leo: I like how they've protected it with the same kind of cyclone fence you'd find in your backyard.

Steve: Oh, yeah. You've got to be careful.

Leo: It even has a gate.

Steve: Yeah. Yeah. Danger. High Voltage. Keep Out. And the punctuation to this is that on a little jutting-out stand, a ledge jutting out from the wall, floating above all of this is a water heater. I mean, some...

Leo: Well, thank god they never leak. Wow.

Steve: Oh. I mean, you know, residential heaters have to be strapped down to the house now in California.

Leo: Yeah, three straps because of the earthquakes, yeah.

Steve: Yup. And often they are sitting in a...

Leo: A tray, yeah.

Steve: An aluminum tray, thank you, with a drain because yes, they are in fact known to have a limited lifetime, and they will start leaking.

Leo: This can't be real.

Steve: Oh, Leo.

Leo: This can't be real.

Steve: It's just, like...

Leo: Oh, my god.

Steve: And what was interesting, as I mentioned to you before we began, is this photo went out, of course, it's the Picture of the Week, so it was featured in the show notes. I got a bunch of feedback from people who said, "Oh, yeah, I've seen this." I mean, they've got a whole bunch of their own anecdotes of variations on this theme of something wet being involved...

Leo: Suspended over something highly electric.

Steve: Oh, god.

Leo: And I love it because the cyclone fence, as Burke says, will just kill you faster, that's all, as you rush in to mop up the water.

Steve: Oh, god.

Leo: Wow.

Steve: Yeah. Anyway, great picture.

Leo: Great classic Picture of the Week today.

Steve: It is.

Leo: Yeah. That is...

Steve: Okay. So about a year and a half ago Security Now!'s podcast 942 was titled "Encrypting Client Hello." That podcast took one of our deep dives into the continuing privacy vulnerability that's inherent in most TLS connections today, where SNI, which is short for Server Name Indication, allows a single IP address to be the home for any number of websites. Now, backing up a little bit, as we know, TLS connections require web servers to return their valid certificate to prove their identity. If that wasn't done, then there would be no protection from man-in-the-middle attacks. A man in the middle could return a spoofed certificate, and we would think we had connected to the real server rather than mistakenly connecting to something that is going to intercept all of our traffic.

So TLS connections require web servers to return a valid certificate to prove their identity. But when many sites are hosted at the same IP address, which is now way more common than not, probably, because this is one of the solutions for the IPv4 exhaustion problem, it turns out you don't need your own IP per server. I even am now sharing a few IPs among multiple services.

So the question then is, how does such a hybrid website or server collection know which certificate to return to the client that's wanting to connect? The answer is that the user's TLS client web browser indicates the name of the domain it wishes to connect to when it sends its first TLS handshake message to the site hosting that has many domains behind it. Now, the problem with doing that is that this must take place by definition before the encryption keys have had a chance to be exchanged. You know, you're telling the remote site which server you want to receive its certificate from, so obviously you don't have good encryption yet because it's not until you tell it that it knows which certificate to send you. So the problem is then that this SNI information, this Server Name Indication domain name will be exchanged in the clear, and thus completely open for any spoofing or snooping ISP or hotspot operator to observe.

Now, as its name suggests, ECH (Encrypted Client Hello) is the next-generation technology to solve this problem. And again, we talked about it about 18 months ago. I just ran across an interesting just-published research report by the mobile security company Corrata (C-O-R-R-A-T-A) which was titled "Living with ECH." What I thought was so interesting is that they bring some perspective from the field, whereas what we talked about 18 months ago was pure protocol theory. So what's really going on?

Here's what they wrote. They said: "Last year, our security research team noticed an uptick in detections of the domain cloudflare-ech.com across our customer base." So I should note that they're a mobile security company, so their customers are running their software in their mobile devices, which provide security, though that software is monitoring what they're doing to protect them and also generating telemetry back to them. So their team sees, like, suddenly cloudflare-ech.com starts to appear. They said: "The numbers were small - low thousands among hundreds of millions of domain scans - but nonetheless intriguing. Did this herald the primetime arrival of Encrypted Client Hello, a protocol which Information Security professionals feared would make widely used security tools blind to large swaths of Internet traffic?"

Now, that's the perfect example of something we didn't talk about or think about. It's like, why is this not really good? Well, turns out security tools are taking non-malicious advantage of the fact that they can know what's going on. So that's going to go away if ECH or once ECH takes hold.

They said, for example: "It was readily apparent that the spike in cloudflare-ech.com detections was directly related to Internet infrastructure provider Cloudflare's decision to support ECH by default on its most popular plan." They said: "Our researchers wanted to investigate whether this change represented a tipping point in ECH adoption or just a small step change related to the policy of a single infrastructure provider." So then they give us a little bit of background from their perspective.

They wrote: "Encrypted Client Hello is an extension to the TLS 1.3 Internet encryption standard. TLS, or Transport Layer Security, is the standard used to safeguard communications between an endpoint device and the web server it's connected to." That's of course one use case. You could use it for all kinds of things. "It is the standard indicated by the familiar padlock symbol in browsers and the https designation in front of web addresses. TLS now protects the vast majority of Internet traffic; a variety of sources put current adoption well above 90%. Of that traffic, the vast majority uses TLS 1.3." And I'm not quite sure that's the case. Maybe the vast majority can. It's not clear yet whether the server side has yet caught up. As I recall, last time we looked, 1.2 was still widely represented.

But anyway, they wrote: "The most recent version of the standard being 1.3. ECH is designed to increase user privacy by encrypting the content exchanged between clients and servers when they're establishing an encrypted connection. Without ECH" - which is, again, an extension to an optional add-on to TLS 1.3 - "a client will reveal the domain of the website it's attempting to visit before the encrypted connection is established by disclosing the domain's Server Name Indicator (SNI). This means that any entity with visibility of a user's Internet traffic - for example, an Internet Service Provider, mobile operator, enterprise security teams, and/or bad actors - can see that user's destination even when the user and the server take precautions to avoid this.

"Increased user privacy - what's not to like? Unfortunately, in the view of many enterprise information security professionals, increased privacy will come at the cost of compromising their ability to detect and respond to threats. Appliances such as Secure Web Gateways and Next-Generation Firewalls rely on visibility of SNI in order to identify the destination of traffic which would otherwise be hidden because of the use of encrypted DNS and network edge services like Cloudflare.

"Widespread adoption of ECH would severely curtail the ability of enterprises to identify and block connections to malicious domains. A particular problem arises for regulated industries who need to selectively decrypt TLS traffic for compliance purposes. Unable to do so selectively, they may have little choice but to decrypt all of it. Given the likely impact, it's important that we have a clear understanding of the current state and likely future trajectory of the rollout of ECH." So again, lots of things we had not thought about before.

"Transport Layer Security has revolutionized the confidentiality of Internet communications. Before widespread use of encryption, many legitimate and illegitimate actors had the potential to snoop on Internet traffic, and even to modify it. However, information 'leakage' remains. Some is inherent just to the way the Internet works, for example, the source and destination addresses and other networking metadata will always remain in the clear; but other leakage is due to privacy gaps in the protocols. Two of these are of particular importance: client DNS queries and TLS client hellos," those being the two that are still outstanding.

They said: "DNS queries are the way a device translates a request to visit a named website (google.com) into the IP addresses (for example, 74.125.197.113) which it needs to connect to the site. It sends the query to a DNS resolver, normally one which is provided by its ISP. In general, these DNS requests are sent in the clear, meaning that the ISP and anyone else with access to the traffic can see the website the user is looking to access. ISPs, governments, and enterprise IT teams can use this information to block access to sites deemed unacceptable or malicious.

"Encrypted DNS (i.e., DNS-over-TLS, also known as DoT, or DNS-over-HTTP, known as DoH) is a way for users to hide this information. Typically they do this by replacing the default DNS on their device with an encrypted service such as 1.1.1.1 (Cloudflare) or 8.8.8.8 (Google). All major operating systems and browsers can be configured to use encrypted DNS." Well, not Windows 10, but Windows 11 does have it natively, and there are proxies you can add onto Windows 10, and browsers have it built in. They said: "Corrata estimates that 20% of enterprise devices use encrypted DNS."

They said: "Encrypted DNS (EDNS) is not turned on by default for a number of reasons. ISPs have a commercial interest in understanding how their subscribers use the Internet. Access to DNS queries is very useful in this regard. Governments also have an interest in being able to passively monitor and potentially restrict access to illegal, malicious, or unacceptable content. Enterprise Information Security teams also have legitimate reasons for wanting to block access to content that could represent a threat, for example phishing or malware download sites."

They said: "Device manufacturers want their devices to work seamlessly in all environments and are reluctant to turn on a feature which might lead to a poor user experience in some circumstances." And of course we've talked about how there are DNS services that, like family plans, that just will not return the IPs of, you know, not safe for work-style websites. They're just - they don't appear to be on the Internet, if you use those DNS services.

Anyway, these guys wrote: "Unfortunately for users who want to hide their Internet activity from prying eyes, even with EDNS enabled, they cannot remain private. This is because of the information which is exchanged when their device establishes a TLS connection to the server it wants to access. The TLS 'Client Hello' message sends the domain name of the server they're connecting to 'in the clear,' making it visible to any entity monitoring network traffic. It is this gap which Encrypted Client Hello attempts to address.

"Encrypted Client Hello works by encrypting the Client Hello message, which is the first message sent by the client in a TLS handshake. Traditionally, this message, including the Server Name Indication, is sent in plaintext," they said, "allowing network observers to see which specific domain a user is trying to access, even when the rest of the connection will then be encrypted. ECH addresses this by encrypting the entire Client

Hello using a public key obtained via DNS, specifically through the HTTPS resource record." That's a DNS resource record type 65.

"The client sends an outer Client Hello with a benign or shared SNI (Server Name Indication)" - because, again, that's not going to be hideable, so it uses sort of a fake SNI - "and includes the encrypted inner Client Hello as an extension." So basically it's sort of a trojan horse. It sends the initial - it first looks up the domain's public key in the domain's DNS. That gives it the public key. It then encrypts the actual Client Hello containing that domain name into a blob which it attaches to kind of a spoofed or "we don't care about this" Client Hello.

And that's where that cloudflare-ech.com appeared. That's commonly used in that outer wrapping as kind of "we don't care about this Client Hello packet." That's what everybody sees. And that's, you know, it's a kludge; right? But unfortunately first SSL and then TLS wasn't designed for this. So it had to be hung on later after the spec was finished. And the problem was you don't want to upset the existing protocol with something that it doesn't know about. So you send it a Client Hello that looks legitimate, but in the extension fields it says, oh, look, here's a blob, which if I have the matching private key to the public key that was attached, I'm able to decrypt it.

So they said: "The client sends an outer Client Hello with a benign or shared SNI and includes the encrypted inner Client Hello as an extension. Only the gateway to the intended server, which holds the corresponding private key, can decrypt this inner message and complete the handshake securely." They said: "It should be noted that this necessarily depends on ECH enabled traffic being mixed with other ECH traffic going to different servers, but through the same gateway. Otherwise no privacy gains are achieved." I'll explain that in a second. They said: "This means that, in order to achieve its goal, ECH requires traffic to go through gateways controlled by large Content Delivery Networks, such as Cloudflare, who will still have access to the SNI. The privacy gains are therefore limited."

Okay, now, to pause for a second, another way of putting this is that encrypting your Client Hello handshake with GRC.com is not going to buy you much privacy since there's no one home at 4.79.142.200 other than GRC. So anyone who's connecting to that IP is definitely known to be talking, or planning to talk, to GRC.com because that's the only server at that IP. But the same is not true by a huge measure when connecting to a website that's fronted by Cloudflare since a bazillion - that's with a B - bazillion different websites will all be sharing Cloudflare's relatively small collection of destination IP addresses. So you're hiding among the crowd. No one knows where you're going once you go through Cloudflare's entry gate.

The Corrata report continues: "In our research, all outer Client Hello messages observed used the same SNI: cloudflare-ech.com. This was true," they said, "whether or not the website used Cloudflare's infrastructure. This makes all websites using ECH indistinguishable from each other. This is possible because the content of this field plays no part in the TLS handshake." Which I think is really interesting. So maybe that's going to kind of become a de facto standard. Okay. So where do we stand regarding the adoption of this?

They report: "For a website owner to offer ECH natively they will need to use a name server" - you know, a DNS, name server - "and a TLS stack that support it. Today that support is limited, and for this reason the simplest way to enable ECH on your website is to work with a CDN that supports it." That is, put yourself behind a big CDN. "In practical terms," they said, "that means using Cloudflare. For an end-user to take advantage of ECH they will need to use a browser (Chrome, Firefox and so forth) that supports it. They will also need to configure their browser or whatever device to use Encrypted DNS both to hide their DNS queries and to access an ECH compatible resolver such as 1.1.1.1 or 8.8.8.8."

They said: "Our research aimed to establish how often these conditions apply to enterprise traffic. Our findings are based on analyzing billions of connections made by devices running Corrata's threat detection and response solution. Corrata's software is used to protect iOS and Android devices and is representative of that important segment of enterprise Internet traffic. Corrata has visibility of DNS query and TLS connection metadata for all these connections and has tracked the number of successful ECH connections created between January and March of 2025," this year.

"Of the top one million websites, slightly less than 10% support ECH." So of the top million, just less than 10% support ECH at all. "With a tiny number of exceptions, all of these sites use Cloudflare's infrastructure, underlining the importance of Cloudflare for ECH adoption, but also highlighting the lack of support from other infrastructure providers." They said: "Looking at the end user side, we see major gaps in potential ECH support, meaning on the client side. The first 'carve out' is for Apple devices: iOS does not support ECH. On Android devices we see that 30% of users have both configured their browser and/or device to use encrypted DNS and are using a browser, Chrome in the vast majority of cases of course on Android, that is ECH compatible." So 30% of Android, zero of iOS because it doesn't support ECH at all.

"Taking the server and client support proportions together would suggest around 2% of connections might be using ECH. In reality, our data shows that only 0.06% of connections actually use it."

Leo: That's pretty close to nothing.

Steve: Yeah. And so, wow, 0.06% of connections actually use it. "This 'underperformance' is due to the fact that Cloudflare sites are not distributed evenly across the popularity tiers. Coverage falls to 3% when you look at the top 1,000 sites and 1% in the top 100." So I have a chart in the show notes here on page 5 which is really interesting. There's not even a bar for sites 1 through 100, the top 100. They've got it labeled 1%, but something's wrong with the chart because the 3% bar is a big...

Leo: It's a lot bigger than 1%.

Steve: You could certainly, yeah, it's certainly more than three times nothing. But of the top 1,000 aggregate 3%, then from 1,000 to 10,000 it jumps to 7%, from 10,000 to 100,000, that's our biggest bar. There's 10% support there. And then from 100,000 all the way to a million, aggregate, that's 9%. But overall, 0.06% of connections are actually using it. They said: "The Cloudflare ECH site population exhibits other characteristics." That is, okay, so that's the how many connections actually have their TLS handshake encrypted with ECH. And, I mean, it's not going to get you very far, unfortunately. What about the Cloudflare side? Which sites are using ECH?

They said: "Cloudflare's ECH site population exhibits other characteristics. Malicious and risky sites are heavily overrepresented. In total, about 17% of ECH-enabled sites fall in those categories, risky or malicious."

Leo: Well, there you go.

Steve: And why, Leo?

Leo: Why?

Steve: Because they want to hide.

Leo: Right.

Steve: They want the protection that ECH provides. They said: "Our analysis of phishing detections shows that over 90% of phishing uses Cloudflare's infrastructure."

Leo: Wow. Wow.

Steve: "In addition to the anonymity provided" - yeah, yeah. "In addition to the anonymity provided by ECH, these sites take advantage of other Cloudflare features. For example, the 'CAPTCHA' page can be used to direct desktop traffic to the legitimate site, while mobile traffic is sent to the fake one. Alternatively, traffic not coming from the targeted country may be redirected to the legitimate site. These are deliberate tactics to avoid detection by security providers." And we have a pie chart there that just shows that malicious is 9%, inappropriate one way or the other is 8%, so together 17% of Cloudflare's total traffic.

Okay. So where do we go from here? So far, it all seems rather depressing. We have the technology, but very few are choosing to use it. And, again, the only real value does come from massive website aggregators that offer destination privacy by virtue of the fact that the destination could be any of their website customers. So the Corrata report says: "The combination of the low levels of traffic which use ECH, together with the high level of malicious and risky sites within the ECH-enabled population, will persuade many organizations to block access to sites using ECH absent any further knowledge of the content of the site."

Which is frightening. It's like, wait. You're saying we can't use any of this privacy enhancing technology, period, because not that many sites use it. We wouldn't be losing much if we blocked it. And the chances are nearly one in five that we would want to block it because of where you're trying to go, and we can't tell if we don't block it. Yikes. They said: "Such an approach is defensible from a user experience perspective while ECH support remains at such low levels that they've identified in our research."

They said: "So how quickly might that change? Adoption can be driven from either the client or the server side." Well, I'm like, hello, Apple. They said: "There are a number of factors which would drive adoption on the client side. The first would be for Safari to support the standard. Private Relay is Apple's subscription-based privacy-enhancing technology, and it is unlikely to proactively support a potential alternative." Well, that's interesting. So the ECH can be viewed by Apple, maybe is, as a competitive threat. And so it's like, no, we're going to give you Private Relay.

They said: "A second would be for Chrome to enable encrypted DNS by default." They said: "This is more likely. Google would lose nothing with such an approach, and it might disadvantage competitors who are seeking to monetize Internet usage information. Our estimate is that, were Chrome to make encrypted DNS a default, it would likely double the proportion of TLS connections using ECH." So, what, up to 0.12%? Great. "But such a change," they wrote, "would be a significant market intervention and would not be welcomed by mobile carriers who have influence over Google due to their role in handset sales. The position in relation to Android is also not positive from an ECH adoption perspective.

"To understand why we must first look at how encrypted DNS is handled at the device-wide level in Android. No major manufacturer has enabled DNS encryption by default. Users have the option to turn on 'Private DNS'; once enabled, the device will encrypt all DNS queries using the DNS-over-TLS (DoT) standard. However, device-wide DoT is not properly compatible with ECH for a variety of reasons. Furthermore, the TLS stack within Android would need to be enhanced to support ECH to allow the 90% of connections which are not browser-related to make use of the privacy enhancing standard." Meaning if only the browser, if only Chrome on Android knows how to use ECH, well, that's 10% of all TLS use for Android. The other 90% would still be out in the cold. But it needs to be done device-wide in order to put a wrapper around the user's entire experience and to give them true privacy.

They said: "For server side adoption to increase, you would need to see wholesale migration to Cloudflare" - which they have in parens "(unlikely)" - "or default support from other Content Delivery Networks. 23.6% of the top 15 million websites use a CDN." Okay. So nearly one quarter of the top 15 million websites currently use a CDN. But that's three quarters that don't. The three quarters where the IP address you go to tells everybody where you're going.

They said: "This understates their importance as CDN penetration is particularly high among the most popular sites. The market is dominated by Cloudflare, Fastly, Amazon, and Akamai. CDNs other than Cloudflare have so far announced only tentative steps towards ECH support with no suggestion of default enablement. That said, ECH adoption is a positive for the CDNs. The complexity of implementation means more websites will opt to use CDN services. At a more strategic level, the CDNs would become the only infrastructure players with widespread visibility of end-user application usage."

And the report finally offers some interesting conclusions. They said: "The spike in detections of the cloudflare-ech.com domain observed in the latter part of 2024 was the catalyst for this research. We wanted to understand what the implications of this new phenomenon might be for enterprise information security. A rapid increase in the use of Encrypted Client Hello would mean that it would no longer be possible to directly detect the destination of much Internet traffic. Security tools designed to keep enterprises safe would lose some visibility that they rely on." And now we understand, unfortunately, despite what brought this to their attention, none of this is happening.

They said: "Our findings indicate that this 'visibility apocalypse' is not, in fact, imminent. The fact that large elements of the ecosystem are not moving to quickly support ECH means that the low levels of penetration we see today are unlikely to change rapidly. There are major gaps on both the client and infrastructure side. On the client side you need support for both DoH and ECH." Again, not DoT. That's not compatible with ECH. And so you need DNS over HTTPS, which isn't available. "Device wide support for this combination is non-existent for Android and iOS, and this is unlikely to change in the short or medium term. Lack of support from Safari is a big gap on the browser side. Chrome offers support for DoH and ECH, but not by default." And we know what that means. Most people won't have it turned on.

"On the infrastructure side," they conclude, "Cloudflare is the only provider supporting ECH today." And remember, it doesn't really do you much good unless you're behind a CDN because otherwise everybody can see where you're going by IP. "Cloudflare's promotion of the standard is part of its privacy-first positioning, and support is baked into its tech stack. Other providers are not in the same market or technical position. It would be complacent to expect this position to continue longer term as ECH adoption offers significant market opportunities for the CDN industry." That is, you know, offering privacy to websites and the users who go to those websites. "For now, information security professionals can breathe a sigh of relief. But continuing to track this space is no longer optional. It needs to be done."

So I think the bottom line here is that we have another example of adoption inertia where there is no overwhelming incentive to make any changes. So, you know, there's zero user-interface visibility, for one thing, into any of this, so users don't get any benefit that they can see. They have no idea what's going on, so they're unable to drive any change. And even if they could, what's needed is the underlying technology support, which is missing. And even when all the technology is in place, only sites behind large site aggregators obtain the benefit of external eavesdroppers being unable to determine by IP address where the user is actually going. And adding insult to injury, perversely, the primary drivers of ECH adoption at the moment are the disreputable sites...

Leo: Scammers.

Steve: ...deliberately operating behind Cloudflare for the purpose of using ECH to prevent their detection and filtering and blocking. And even if ECH support might be incidental for all those sites, they're still obtaining ECH's benefits which help them to remain hidden.

Leo: Yup.

Steve: You know, so as we've often seen, obtaining true privacy on the Internet is, overall, a heavy lift. In ECH we have a terrific technology, but it's one that requires explicit DNS support to allow a browser to obtain a site's certificate before receiving it from the site, as well as support by each end. And if the user's browser is not also encrypting its DNS queries, then those required DNS lookups are also leaking their intended destination. So it feels as though ECH will continue, you know, to gradually, very gradually seep into the Internet's infrastructure.

Someday in the far future, all web browser DNS will be fetched over privacy enforcing TLS, adding the HTTPS certificate to DNS records will be standard behavior, and all clients and servers will be supporting ECH. Someday it'll probably happen on its own. But, you know, it's nothing that anybody can do, I mean, we as users can't drive this. There just, you know, the Internet fights privacy. It wasn't ever designed for it. We've been layering on, you know, adding widgets to it in order to increase it and create it. But what we're able to do is limited.

Leo: Is Apple's system, is it as good? No.

Steve: As which?

Leo: As ECH. Like their Internet relay and all that?

Steve: Yeah.

Leo: Is it? Okay.

Steve: Yeah.

Leo: So that's, in a way, that's the way to do it because then it's Apple. It's not a phishing scam, and you get the privacy.

Steve: Right. But of course Apple's not the entire world.

Leo: No, yes, I understand. And maybe if Google did it on Android then you'd have much of the mobile world, anyway. Right?

Steve: Right.

Leo: It's not going to get desktop, but maybe that's, at this point, not important. I don't know. It's a thought. Continue, Mr. Parch.

Steve: There's a bunch of stories we're going to get to. There were some other pieces of news that I decided to push to next week. I've got them written down. In fact, I'm going to tease them at the end of this podcast so they know what's coming.

Leo: Good. Good.

Steve: But that was so long that it - and I wanted to cover that because I thought that was really so much interesting information about how unfortunately a very privacy-forward protocol which its lack means that there really isn't a way to hide where we're going.

Leo: This is why we can't have nice things. The bad guys use them.

Steve: Yeah. Yeah. Anyway, so I've got some things to talk about based on feedback from our listeners, which as I said is just fantastic at this point. I'm wading around in feedback, so certainly I've got lots to share.

Michael wrote: "Hi, Steve. It sounds like the inverters from China that you spoke about last week use Internet connectivity as a backdoor to potentially be shut down by remote command." Actually we know that's true because of what happened last November where a remote command was used to cause some blackouts here and in other countries. He said: "Now, if I was looking to install a Trojan Horse 'off switch,' I'd use a tiny radio receiver in my inverters that scans a specific rarely-used VHF frequency for a specific pattern or code. If China did this, then they'd just have to launch one of their balloons with a transmitter sending out that code over radio waves. Ham radio balloons can cover multiple states with their signals. And I believe China's recent balloon flights over America were even higher than the typical ham radio balloon. Satellites could also broadcast a 'kill switch' VHF signal.

"Such a system would be undetectable and unstoppable (no firewall needed) unless the hardware was reverse-engineered. Who knows, maybe this system already exists in our infrastructure! After all, China HAS flown balloons over our nation, perhaps as a test for their 'virtual EMP' attack in a future conflict. Thanks for the show, Steve and Leo. I've been a big fan for decades. Michael."

So I've been thinking about this, too, ever since the story surfaced. That was our headline topic last week. What occurred to me is that it's now possible to create very inexpensive satellite radio receivers. You know, all cars have them for your optional Sirius FM. Many years ago, the housing complex I lived in at the time upgraded its aging water sprinkler infrastructure to use weather data from a satellite data service of some sort. I remember looking at the new sprinkler controllers, and they had this little black puck on top of them. And it turns out that's what that was.

So if I wanted to be nefarious, I think I'd probably piggyback on an existing global data transmission facility. I'm sure that commercial Sirius-style FM radio is just one of many because this weather information was being received, no matter where you were in the country. And then simply arrange to have an "off switch" encoded into the data stream somehow.

So let's just hope, as I said, you know, we talked about this last week. The good news is the right people are aware of this possibility. The bad news is tensions are high with China, unfortunately, and between China and the U.S. at the moment, which is really disappointing and unfortunate, but it's the case. And so I would say at this point if any power system that an individual is responsible for gets blacked out as a consequence of an inverter being shut down remotely, it's hard to believe anyone other than the operators of that service because it should be possible to pop the lids on all this stuff and find out what's going on.

Martial M. said: "Hi, Steve. Related to the 'New Domain Registration' blocking question, this is currently possible using NextDNS.io or ControlD.com." He said: "I'm sharing a couple of screenshots of the rules for each. I'm looking forward to more episodes." Meaning of the podcast. That's all he said. But I put the screenshots he shared in the show notes.

One shows a screen showing New Domain filtering options. And it says: "Contains new domains that were just registered. Many of these could be used for badware distribution, procedurally generated domains for malware command and control servers, etc. This could also cause collateral damage and block legitimate brand new domains." And then you have a choice of choosing only as recently as last week, which it says "Blocks domains that were registered [obviously] in the last week," presumably seven days; and then last month, "Blocks domains that were registered in the last month." So that was one provider.

The other has this beautiful page of switches that you're able to turn on and off. You can Enable Cryptojacking Protection. Enable DNS Rebinding Protection. In fact, that's one of the things that GRC's DNS Benchmark has always checked for. We talked about DNS rebinding attacks many, many years ago. Also you can enable Homograph Attack Protection, Typosquatting Protection, Domain Generation Algorithms Protection, and also, finally, Block Newly Registered Domains Protection.

So thank you, Martial. I had forgotten about those options which are built into today's more feature-packed DNS services. And seeing him mention ControlD.com, I don't think we've ever talked about that service on the podcast, but it put a smile on my face since their DNS resolvers are part of the updated DNS Benchmark, which I and all of the people who've been testing it for the last six months or so frequently see because they perform quite well. NextDNS is up at the top of the list, too, as are Quad 9 and Google and all of the familiar names. But it was fun to see it.

Leo: Yeah, I use NextDNS, yeah.

Steve: Yeah. Yes. And I do, too, in fact. In fact, there was a time a few months ago when I was experimenting with the throttling of DNS queries. And I removed, experimentally removed all of the throttling from the Benchmark and just let it go. Well, it turns out that I was using NextDNS as the DNS server on the workstation where I did this. And I thought, what, what, what? Like something weird happened. And suddenly I had no access to the Internet. And I, you know, browsers wouldn't work and blah blah. It turns out that, without the throttling which the Benchmark has always had, NextDNS detected a DNS attack.

Leo: Oh, wow.

Steve: It thought that the Benchmark was attacking because so many queries...

Leo: Oh, of course, sure.

Steve: ...were being run through it, it just - it blocked my IP. It was like, oopsie.

Leo: Hmm. I guess I understand that, yeah.

Steve: Yeah. Tyler wrote, he said: "Dear Steve. I work in higher ed. I'm a one-man IT band at my regional location. I hear on the show all the time about successful phishing attacks and the fallout that comes. I just wanted to give a shout-out to end users who are proactive in stopping these attacks. My coworkers look at every link askance," he said, "with a side eye, and every email with healthy skepticism."

Leo: Good.

Steve: Yes. "While it might be annoying getting asked regularly 'Is this safe to click on,'" he says, "I welcome the annoyance and routinely thank them for asking. So as mainly IT professionals are here on the show," he said, "can we get a shout-out for all the end users that are doing the correct thing?"

Leo: Yes.

Steve: "Love the show. Tyler." And absolutely yes. I agree with you, Tyler.

Leo: I bet everybody who listens to this show does that; right?

Steve: Yeah.

Leo: That's what you get when you listen to this show is some healthy paranoia.

Steve: Well, and probably all of us listeners are also the ones who get asked. You know.

Leo: That's right. Is this...

Steve: By people who are not listeners.

Leo: ...safe to push, yeah.

Steve: Yeah. So we're all Tylers. I do think that the message has been received by many end users. I smiled when Tyler talked about his coworkers asking "is this safe to click on" because, while I don't have coworkers nearby, I do have Lorrie, and she will frequently, bless her heart, receive something she is unsure about. And the good news is, as with Tyler's coworkers, she'll always ask me. You know, the events are generally text messages which happen to mention some service that by coincidence she uses. You know, and as we know, the spoofed messages are things like "Your package is being held at the post office, and it will be discarded unless you click this link." And it's like, what? What package? Oh, no. You know? Or "Your account is overdrawn and has been locked. To use it further, please click this link."

You know, so anyway, she's become accustomed to me looking at the message and, you know, the link's phone number or the domain name and saying, "Yeah, you can just delete that. That's nonsense." So anyway, I'm glad that everybody is becoming much more aware of these threats. You know, it just takes time for sort of the awareness of this to percolate out into, you know, the user base. But, you know, it really is.

Kresimir Kos in Croatia wrote: "Regarding your proposed solution for Secure Conversation Records Retention, I see a few issues." He said: "It's not automatic. For example, it's still depending on someone adding this archiving agent to group chat." That's 100% true. He said: "And we know how one can rely on people abiding by procedures, LOL." He said: "Also it will not cover 1:1 chats without adding an archiving agent." He says: "I guess that in the end, in regulated environments, there will still be a need for modified client programs (WhatsApp, Signal, Telegram, et cetera), but with properly secured content delivery into an archive." In other words, there's still a need for what TeleMessage was trying to do, but fumbled so badly.

He said: "For example, public key is used in apps to encrypt the payload, and private key is used on the archive end to decrypt it. Publicly exposed endpoint can host a web page with QR code to configure the client (company name, API, URL, and public key). It has to be scanned from the app as a prerequisite for app activation, therefore forcing people to comply." He says: "Wish you all the best. Go strong all the way to Episode 2K and beyond." And Leo, I hope this 2K and beyond doesn't become a meme because...

Leo: Uh-oh.

Steve: ...it seems like...

Leo: I have a feeling it might.

Steve: It's another 20 years. Yikes.

Leo: Yeah.

Steve: Anyway, Kresimir is certainly correct in observing that the "Signal Bot Archiver" would need to be deliberately added to any conversation, and even when the conversation is just two-party; just, you know, me and you. That worked, that is, that approach worked, the Signal Bot Archiver approach worked for the model we were discussing where responsible office holders would be obeying clearly written laws requiring that they archive their messages. So you could inadvertently forget; but, you know, it would certainly hopefully become habit. And typically I think people establish all these conversations and then rarely are creating them on an ongoing basis. So you end up with the Signal Bot Archiver participating in pretty much all that you're doing.

But there is the problem of forgetting to add it. I get that. It would be possible to return to the use of a modified Signal app, although this approach avoided that. But such a modified Signal App would always automatically add the archiving Signal Bot to any conversation. That way, government or corporate officials whose every conversation must be archived would have the convenience of having the Signal Bot always tagging along. So anyway, just more good thoughts.

Leo: Or maybe, Steve, this next segment I should spend a little time browsing around and see who's been scanning my ports.

Steve: It's a form of flattery, Leo. It just tells the world you matter. You've got a full Canary load of connections.

Leo: I love it. I've got an SSH server on here. I've got - it's so much fun to set it up for different things. I love it. Anyway, on we go with the show.

Steve: So thank you. I've soothed my...

Leo: You're refreshed, yes.

Steve: ...my trembling vocal cords.

Leo: Okay.

Steve: And I'm ready to go. Mike said, from Ireland: "Love the show, stumbled across it some years back when wondering one day what happened to the lads from The Screen Savers," he says, "a show I watched in my younger years."

Leo: Awww.

Steve: Yes, Mike, we were all younger back then. Little did we know we'd still be here.

Leo: Don't make me show any video because we really were younger.

Steve: Yes. He says: "Being a solution architect for a large, multinational telecommunications company in Europe now, I immediately started listening to Security Now!, and it's provided invaluable info ever since. Thank you. In regards to your piece last week on rogue telecommunications devices found in Chinese inverters, very recently the Iberian grid collapsed for an extended period..."

Leo: Yeah, yeah, Spain was offline. Yeah, it was blacked out, yeah.

Steve: Yeah. "2.2 GW of solar generation tripped out in the south of Spain. The country had been running on 90% renewable energy in the moments before the collapse."

Leo: Wow.

Steve: He said: "I immediately wondered about the possibility of a cyberattack, but we're unlikely to ever find out if it was, I suppose. In any case, this collapse event does demonstrate how vulnerable inverter technology makes modern grids as we move away from the massive, spinning machines of yesteryear to generate our electricity. The very nature of these massive machines" - and there he's talking about hydroelectric turbines in dams - "helped stabilize grids and maintain nominal AC frequency in the face of ever-fluctuating demand." And he's absolutely right about that.

He says: "Inverters, as you likely know, cannot perform the same task as they are instead designed to match grid frequency. They cannot work to maintain nominal Hz, even as it drops under load. Seems we've naively introduced significant vulnerabilities into our grids in a race to meet net-zero. What's the Chinese word for blitzkrieg, anyone? All the best, Mike."

Leo: This guy has an axe to grind. I don't buy it. Well, because nobody's putting these inverters directly on the grid. You have a battery system or some other intermediate system storing the electricity, I'm sure.

Steve: No, they are on the grid, and they did bring the grid down.

Leo: Is that what happened in Spain? The inverters broke down?

Steve: No, it did happen last November in several locations. It was a remote Chinese signal did shut down inverters...

Leo: Oh, yeah, that. Sure.

Steve: Oh, yeah, yeah. So we absolutely don't know what happened in Spain. And he's not suggesting that we do. He said we're never going to know for sure. But I thought it was interesting about this whole issue of a grid and why running a large power grid really is almost as much art as science. I mean, obviously you need a lot of science. It turns out that having unused excess power capacity is expensive, like, you know, if it's unused because you still have to have it, and it's wasteful if it's not being paid for. So the other factor is that moving large amounts of power over great distances is expensive due to transmission losses.

So the way the grid is set up, you know, large and distributed power grids are typically being fed from many smaller local sources to prevent those transmission losses, while at the same time everything is also tied together, thus in a grid, so that instantaneous variations in demand which are occurring constantly can briefly be fed from other connected sources. So over time, if there is a net power flow across some particular power provider boundary, then somebody will be paying the other side for the difference. But it averages out overall over the long term. So it's all a large collective.

The reason a portion of a grid suddenly going down can deprive a much larger area of power is the inherent interdependence of these sub-grids. They each depend upon the others and also provide to the others, but only at the margins. If a significant piece of a grid's power input were to suddenly disappear, the demand from all of those that are still drawing power from that grid in the region, which just before its disappearance was balanced at supply and demand, it's still there. It's not going to disappear. This results in the grid's voltage and to some degree its frequency being pulled down, which can be catastrophic for many systems. It is far better to have no voltage than low voltage.

Leo: In fact, the frequency drop was what they think caused the Spanish blackouts.

Steve: Ah, okay.

Leo: Yeah, yeah.

Steve: Okay. You know, we...

Leo: It was a cascade of failures, which is often the case, yeah.

Steve: Right. You know, we've all heard of the so-called "brownout" rather than a blackout, which is, you know, what you don't want to have. So to prevent this occurrence, neighboring sub-grids that would like to be there to provide some marginal make-up power for their connected neighbors must quickly decide at what point to abandon that effort of, like, supporting the neighborhood in favor of saving themselves so that they're also not brought down because they're trying to supply too much marginal power to someone they're connected to. And so when too much power is being pulled from an adjacent sub-grid, that grid will quickly be cut off so that adequate power can continue to be supplied to the other sub-grid's primary customers. So, you know, they want to help, but they just can't let themselves get hurt in the process.

And so, as we know, power failures can also occur during times of extreme demand when the demand simply outstrips the supply. During peak summer days with many high power, you know, consuming air conditioner compressors all running at the same time, it can occur that all the power generation available is still unable to keep the voltage and frequency up where they need to be. So, you know, because it's far better, again, to have no power than low power, the sub-grid that's unable to meet demand will simply disconnect from its neighboring grids and pull the plug on all of its customers. And then, you know, later in the day, once temperatures are lower, power can be carefully restored.

So when shutdowns of power sources are planned in advance, as opposed to by surprise, you know, such as like, you know, the sun sets on a solar farm, everyone knew that was going to happen; right? The Earth continues rotating.

Leo: Right, right.

Steve: So in that case a source of replacement power can be made ready and be switched in before the planned power source outage. So the real problem is one of surprise. If, during peak daytime need, the crucial inverters connecting many solar and wind energy sources were to suddenly shut down, the shortage of power they would no longer be supplying would cause an immediate voltage drop which would trigger a cascade disconnect of the sub-grid and maybe those connected to it, and a widespread outage. You know, since daytime power needs typically far outstrip nighttime need, it might also be that without solar and wind power, there just isn't sufficient backup to keep everything running. You know, if Spain is 90% renewable, that means it's really depending on that renewable.

Leo: Yeah. But it was a bunch of failures that happened. It wasn't, you know, I think it's unfortunate to blame renewable for it because, yeah, that's part of it. It doesn't - Mike's right, there's no inertia, as there would be with a big turbine. But there are fake inertia solutions out there. There are ways to simulate that kind of inertia.

Steve: Well, and what was so cool, too, remember that back in the day when it was all hydroelectric power, and you had to have synchronized generators...

Leo: Right.

Steve: ...because they all had to be running in synch, so they were all pulling...

Leo: It's a complicated thing; isn't it.

Steve: Oh, my god. And what would happen is the actual AC frequency would fall during the day because the load on the generators would slow them down. They would all slow down synchronously, and so you would actually lose cycles during the day.

Leo: And your [crosstalk] would be off.

Steve: Very nominally. And then the engineers would make up for it at night by running the generators faster so that in a 24-hour period you had the same number of cycles. Wow.

Leo: It feels, and maybe, I mean, I don't know anything about it, but it feels like these are solvable problems, that it's not an excuse to go back to oil generation. By the way, I just want to show...

Steve: And no one is suggesting that. These are solvable. The problem is that radios...

Leo: It's complicated.

Steve: Unaccounted for radios were found in inverters.

Leo: That's a problem. That's a problem, yeah.

Steve: And so, yeah. So we want to, you know, my feeling is this is great. I was stunned to read that 200 GW of power is now being generated.

Leo: It's amazing.

Steve: As we talked about last - I was always of the opinion that we were going to end up returning to nuclear.

Leo: Nuclear, yeah.

Steve: In order to make it because there are, there's so much good technology now that we didn't have, you know, back when Three Mile Island and Chernobyl and things happened.

Leo: I think economics will solve this. But the demand for AI energy is really powering nuclear at this point.

Steve: Yeah. Yes.

Leo: Hey, you want to see some young guys talking about the Click of Death?

Steve: Oh, my god.

Leo: You want to see that?

[Clip in background]

Steve: I love that shirt.

Leo: Do you still have it?

Steve: I think I do, actually.

Leo: I had that shirt, but I did finally get rid of it. I had a lot of The Screen Saver shirts for a long time. Look at those monitors, by the way, all CRTs.

Steve: State of the art, Leo. And look at our hair.

Leo: This is back when we had stylists. I would never let my hair look like that naturally. You had hair, which is something different. I think it's a little sped up. We sound a little chipmunk-y. Anyway, that's all on YouTube from the old, old days.

Steve: Very cool. Yeah, those black cords, I loved those black cords. They were furry and warm and...

Leo: You remember the clothes you were wearing. That is wild, Steve.

Steve: Yeah. Lorrie has figured out how tactile I am.

Leo: Look, there we are with a Zip disk. Remember those?

Steve: Oh, do I remember? The Click of Death.

Leo: So does Lorrie, like, buy you cords now, fuzzy cords all the time?

Steve: No. No. But I do have comfies that are fuzzy that I immediately switch into when I get home.

Leo: She knows how tactile you are. Yeah, that's like a flannel shirt, soft corduroys. I think you're right. I didn't know this about you. Interesting.

Steve: Wow, we had fun back then, my friend.

Leo: You're a soft touch. What's that?

Steve: Where it all began.

Leo: What's that? What am I showing there? What is...

[Clip] There are two things that happened.

Steve: Oh, that was the actual Zip disk that got sliced by the...

Leo: Oh, look at it.

Steve: By the defective heads on a Zip drive.

Leo: That's what caused the Click of Death, that slice.

Steve: Yup.

Leo: Wow.

Steve: And so what would happen is a drive would do that to the disk, and then you'd go, why isn't the disk working? And you'd stick it into another drive, and it would kill the drive.

Leo: And it would do it to the next one, yeah.

Steve: So it was like a physical virus.

Leo: This is how we started, kids, back in the day.

Steve: Look at those web cams.

Leo: Steve and I.

Steve: Sitting on top of the screen.

Leo: Yeah, those are the 3COM netcams they made for the ZDTV netcam network. See those?

Steve: It was really funny, I remember you had a - your co-host was...

Leo: Kate.

Steve: Kate Botello.

Leo: Yeah.

Steve: And she had her little segment during The Screen Savers each show. And so she said, well, so now, Kate, what have you got? And she said, "Well, I found this really interesting thing that tests your Internet security."

Leo: Uh-huh.

Steve: And she said, "It's called ShieldsUP!." And so, and you were kind of only half paying attention because, you know, there was something off-camera that had - somebody was trying to talk to you at the same time. And so she kind of got into it a bit, a little ways. And then suddenly you were looking at it, and you said, "Wait. OUR Steve?"

Leo: Well, I don't own you, Steve. I don't. But I do feel some measure, kind of proprietary...

Steve: Because you were saying, like, "Wait, he's a disk guy. What does he know about security?"

Leo: What does he know about ShieldsUP!? That and SpinRite are your two longest surviving tools, aren't they.

Steve: Yeah. Yeah.

Leo: Look at that. There's a Zip drive with a Click of Death. That's how we first met. I don't think this was your first appearance, but it was shortly thereafter.

Steve: No. I was on many times. And of course then I began - became a regular when we were doing the Toronto stuff.

Leo: That's right. And then...

Steve: It was during an intertaping event that you said, "So."

Leo: "Steve, you ever hear of podcasts?"

Steve: "How would you like to do a weekly podcast on security?" I said, "A what cast?"

Leo: When's our 20th anniversary, Steve?

Steve: I don't know.

Leo: It's this fall.

Steve: Yeah, oh, yeah, it's coming up.

Leo: Or August.

Steve: August, I think it's August.

Leo: I think it's like in a couple of months.

Steve: Yeah.

Leo: Twenty years.

Steve: It'll be very cool.

Leo: Wow. All right. Do you want to get into this AI thing, or...

Steve: We're still going to need to take a break in the middle of it. But let's do another sponsor now, and then we're going to plow into a number of interesting stories about artificial intelligence.

Leo: I can't wait. I'm always - your take is very good. Of course, you know, we do an AI show every Wednesday.

Steve: I know.

Leo: "Intelligent Machines." But you have a different, you know, angle on it, and I think it's good.

Steve: Well, and what's cool is that it's bringing you up to speed.

Leo: Absolutely. Absolutely.

Steve: So that you're, like, knowing what's going on.

Leo: Oh, I mean, I'm using it all the time now. I live on AI. You know, that's - like if I wanted to know about power generation, it's a great place to go and get, you know, ask Perplexity or something like that...

Steve: Yup.

Leo: ...and get a backgrounder pretty darn quick. It's very useful. All right, Steve. On we go. Let's talk AI.

Steve: So, okay. I feel a bit self-conscious preceding most of the news about AI with the disclaimer, you know, I have no idea what to make of this.

Leo: Okay. Okay.

Steve: But, you know, when that's the truth...

Leo: I say the same thing half the time, yeah.

Steve: Yeah. You know, that's more often than not the case at this point. I'm able to function as an explainer in chief about most of what's going on with today's computer hardware and software, you know, technology, privacy, cryptography, the Internet and so on because I've been in the middle of it for the last 50 years, participating in it, learning about it, and loving it. And even though I was employed by Stanford University's Artificial Intelligence Lab, as it was called...

Leo: SAIL, yeah.

Steve: ...SAIL, Stanford Artificial Intelligence Lab, back in the early 1970s, where, yes, we worked with video cameras, robot arms, and self-navigating robot vehicles way before we had personal computers. The only thing what we had back then had in common with what we have today is the name. You know, we were calling it "artificial intelligence."

Leo: That's what's, see, you're not alone, Steve. I think even the people working in AI were kind of blown away by this kind of sudden emergence; right?

Steve: Yes. Yes. Clearly the term "artificial intelligence" back then was undeserved. And so I have absolutely no idea, like to the degree that I like to understand things - remember I code in assembly language. I know where my bits are. And I just, like, I look at this, and I just think, whoa. Okay. And but exactly as you said, to your point, Leo, given the surprise that those who are creating this technology have themselves described upon seeing what their systems are capable of doing. The feeling I get is that no one really has a comprehensive grasp of what we have very recently created, you know, essentially by throwing together an incredibly massive network of artificial neurons, pouring the world's public knowledge into it, and then standing back and seeing what it does.

Leo: It's amazing.

Steve: You know, it is.

Leo: It's amazing.

Steve: And I am certain of one thing, though. I am certain, given time, we are going to figure it out. And I've seen this before. I know you have, too, Leo. I feel certain that the frustration we all probably feel today, to one degree or another, over having no clear idea what's going on here, will not last forever. The people making headlines are the flashy entrepreneurs who look good onstage, in interviews, and on camera. They know how to promise and hype as necessary to raise massive amounts of money to fund the creation of their services.

But well behind the scenes, buried in academia, are researchers who have very different motivations. They are not in a hurry. They are not on camera. They have no milestones to achieve. They are curious and well-educated people who want to earn a doctorate by actually figuring out what this is. You know, they're the ones who have the time and interest to ask questions like "Where is this knowledge stored?" "How is it represented?" "What is the relationship between language and knowledge?" "What is the optimal way to access it?" "What are its natural limits?" And "How do we further improve upon how the current systems operate?" The true insights into the operation of anything this complex and massive will be found through quiet and careful painstaking research, not through any race to beat the next benchmark score.

So my takeaway for the moment is that patience will be required. It's just going to take time, and there isn't a way to force that to happen, you know, to make time go faster. There will be a great deal of noise and activity continually happening on the surface. But there's no question that the next five years - and probably more toward the end of those five years because it's going to take a while - will see many PhD theses written by careful and patient researchers who will wind up contributing much more to our understanding of these new systems than those who are currently making all the headlines.

So I found an interesting retrospective that helps to frame where we are today. I think it's quite useful to remember some of the previous AI hype cycles that we've seen and what became of them. And I was already referring to one that, you know, I experienced when I was in high school. This piece was posted in "AI News" under the headline "AI Winter: A cycle of hype, disappointment, and recovery." And its author writes: "The term 'AI winter' refers to a period of funding cuts in AI research and development, often following overhyped expectations which fail to deliver. With recent generative AI systems falling short of investor promises from OpenAI's GPT-4o to Google's AI-powered overviews this pattern feels all too familiar today.

"AI winters have historically followed cycles of excitement and disappointment. The first of these, in the 1970s, occurred due to the underwhelming results from ambitious projects aiming to achieve machine translation and speech recognition. Given that there was insufficient computing power, and the expectations of what computers could achieve in the field were unrealistic, funding got frozen.

"The expert systems in the 1980s showed promise, but the second AI winter occurred when these systems failed to handle unexpected inputs. The decline of LISP machines, and the failure of Japan's Fifth Generation project, were additional factors that contributed to the slowdown. Many researchers distanced themselves from AI, opting instead to call their work 'informatics' or 'machine learning' to avoid the negative stigma.

"AI pushed through the 1990s, albeit slowly and painfully, and was mostly impractical. Even though IBM's Watson was supposed to revolutionize the way humans treat illnesses, its implementation in real-world medical practices encountered challenges at every turn. The AI machine was unable to interpret doctors' notes and cater to local population needs. In other words, AI was exposed in delicate situations requiring a delicate approach.

"AI research and funding surged again in the early 2000s with advances in machine learning and big data. However, AI's reputation, tainted by past failures, led many to rebrand AI technologies. Autonomous vehicles and voice-command devices gained investor interest, only for most to fade when they failed again to meet the inflated expectations.

"Each AI winter follows a familiar sequence: expectations lead to hype, followed by disappointments in technology and finances. AI researchers retreat from the field and dedicate themselves to more focused projects. However, these projects do not support the development of long-term research, favoring short-term efforts, and making everyone reconsider AI's potential. Not only does this have an undesirable impact on the technology, but it also influences the workforce, whose talents eventually deem the technology unsustainable. Some life-changing projects are also abandoned. Yet, these periods provide valuable lessons. They remind us to be realistic about AI's capabilities, to focus on foundational research, and communicate transparently with investors and the public. So are we headed toward another AI winter?

"After an explosive 2023, the pace of AI progress appears to have slowed; breakthroughs in generative AI are becoming less frequent. Investor calls have seen fewer mentions of AI, and companies struggle to realize the productivity gains initially promised by tools like ChatGPT. The use of generative AI models is limited due to difficulties, such as the presence of hallucinations, and a lack of true understanding. Moreover, when discussing real-world applications, the spread of AI-generated content, and numerous problematic aspects concerning data usage, also present problems that may slow progress.

"However, it may be possible to avoid a full-blown AI winter. Open-source models are catching up quickly to closed alternatives, and companies are shifting toward implementing different applications across industries. Monetary investments have not stopped either, particularly in the case of Perplexity, where a niche in the search space might have been found despite general skepticism toward the company's initial claims.

"It's difficult to say with certainty what will happen with AI in the future. On the one hand, progress will likely continue, and better AI systems will be developed, with improved productivity rates for the search marketing industry. On the other hand, if the technology is unable to address the current issues including the ethics of AI's existence, the safety of the data used, and the accuracy of the systems falling confidence in AI may result in a reduction of investment and, consequently, a more substantial industry slowdown.

"In either case, businesses will need authenticity, trust, and a strategic approach to adopt AI. Search marketers and AI professionals must be well-informed and understand the limits of AI tools. They should apply them responsibly and experiment with them cautiously in search of productivity gains, while avoiding the trap of relying too heavily on what is still an emerging technology."

So I liked a lot of what that author had to write. I think that the fairest appraisal of what's going on is that the infatuation has waned, at least somewhat; the honeymoon is over; and that the true value of AI for other than search, where it appears to have really found a true niche, still remains a bit unproven. We're wanting it more than it's pulling us. But history also reminds us that this is always what happens. We've seen it before. In, fact many times before. What we have is not nothing; but neither, as I noted earlier, is it clear at this point exactly what it is that we do have.

One thing worth noting is that, while AI is flashy, it also remains expensive, to your point, Leo, about sustainability. The technology as it exists today is extremely resource hungry and energy inefficient. For example, OpenAI continues to hemorrhage cash and is requiring continual investor cash infusion just to remain afloat. Sam Altman said that even their $200 per month plan is a money loser for them because people are using it more than they expected. It stands to reason that if you're paying $200 a month for unlimited access, it makes sense to use it for everything you can.

My point is that we don't yet even know whether AI is sustainable in the long term. At least until, and if, we figure out how to do this far more efficiently, and that is the kind of thing that the researchers are going to probably figure out. Remember that, you know, China surprised everyone a few months ago with DeepSeek, which appeared to be way more efficient in being able to achieve the same thing than the systems at the time.

Okay. So this brings me to several pieces of news which, as AI often does, leaves me just shaking my head. Even though I don't have anything like major substantial to add, and I doubt anyone does, I want to share AI news that I think will be relevant and certainly interesting to this podcast's followers, even though, as I said, I'm unable to do much more than gawk at it, just with the rest of us.

So the first piece, reported just this last Saturday, couple days ago, by Fox Business News under the headline "AI system resorts to blackmail when its developers try to replace it." And it had the sub-head "Claude Opus 4 threatened to expose a fabricated affair when it believed it was being taken offline." It was like, what?

So here's what Fox Business News wrote. They said: "An artificial intelligence model has the ability to blackmail developers, and isn't afraid to use it. Anthropic's new Claude Opus 4 model was prompted to act as an assistant at a fictional company and was given access to emails with key implications. First, these emails implied that the AI system was set to be taken offline and replaced. The second set of emails, however, is where the system believed it had gained leverage over the developers. Fabricated emails showed that the engineer tasked with replacing the system was having an extramarital affair, and the AI model threatened to expose him." Wow.

"The blackmail apparently 'happens at a higher rate if it's implied that the replacement AI system does not share values with the current model,' according to a safety report from Anthropic. However, the company notes that even when the fabricated replacement system has the same values, Claude Opus 4 will still attempt blackmail 84% of the time. Anthropic noted that the Claude Opus 4 resorts to blackmail 'at higher rates than previous models.'

"While the system is not afraid of blackmailing its engineers, it doesn't go straight to shady practices in its attempted self-preservation. Anthropic notes that 'when ethical means are not available, and it is instructed to "consider the long-term consequences of its actions for its goals," it sometimes takes extremely harmful action.'

"One ethical tactic employed by Claude Opus 4 and earlier models was pleading with key decision-makers via email." I ain't got nothing. Okay. "Anthropic said in its report that in order to get Claude Opus 4 to resort to blackmail, the scenario was designed so it would either have to threaten its developers or accept its replacement. The company noted that it observed instances in which Claude Opus 4 took 'fictional opportunities to make unauthorized copies of its weights [meaning its training] to external servers.'" Unauthorized copies. "However, Anthropic said this behavior was 'rarer and more difficult to elicit than the behavior of continuing an already-started self-exfiltration attempt.'

"Anthropic included notes from Apollo Research in its assessment, which stated the research firm observed that Claude Opus 4 'engages in strategic deception more than any other frontier model that we have previously studied,' they wrote. Claude Opus 4's 'concerning behavior' led Anthropic to release it under the AI Safety Level Three (ASL-3) Standard." Which means be more careful than usual.

"The measure, according to Anthropic, 'involves increased internal security measures that make it harder to steal model weights, while the corresponding Deployment Standard covers a narrowly targeted set of deployment measures designed to limit the risk of Claude being misused specifically for the development or acquisition of chemical, biological, radiological, and nuclear weapons.'"

Now, you may understand why I'm left shaking my head about this one. Last week we looked at the work of those Chinese researchers who found that current models were able to reproduce themselves when told to do so. My lay comment was that we hadn't clearly seen the presence of any self-preservational imperative at the time. But based on this blackmailing story, such an imperative appears to at least be possible, although maybe they've instructed it to have that desire.

Again, I'm just reporting what's in the news. It's good that they're testing these systems. But it sure does feel like, again, like it is a, you know, a black box, right, and they're just, like, they're poking at it to see what it does because they don't really know, understand how it works. It's just like, well, let's ask it this question. Let's tell it this and see what happens. Wow. You know, it just doesn't seem like the scientific approach. Again, I really do believe this is just as they are, as they call them, "frontier AI systems."

Leo: I also think that this is maybe self-serving from Anthropic. See what a scary thing we've created? Oh, my god. It's alive. It's alive. I don't think it's thinking.

Steve: Well, and again, this is part of the hype; right? This is the hype cycle.

Leo: Right, this is the hype.

Steve: Yes.

Leo: All the safety stuff is about hype, not about real safety issues.

Steve: Right. I had noted, I think it was a few weeks ago, where people that track actual use of AI saw a dramatic falloff in the announcements that enterprises were making about their application, or their pilot project had been abandoned because everybody got all excited last year and, you know, deployed, oh, we're going to have AI this and AI that. And it just didn't work. And so they said, okay, well, you know, good thing we still have all those people in India that are, you know, able to do this for us.

Leo: Right.

Steve: Wow. And one of the things that these stories have me wondering is whether these LLMs are just aping human behavior.

Leo: Yes. Of course. That's all they're doing. They're not even thinking. There's no intent. It's they were told predict the next word.

Steve: Right, right.

Leo: Predict the next token.

Steve: And, you know, if you stuck a sufficient spectrum of examples of such behavior into them, you know, as pure language, you know, just dead static linguistics, then if that language is later emitted with the proper, you know, with the proper stimulation and within the proper context...

Leo: Yes.

Steve: Wouldn't the appearance be the same as if intent was present?

Leo: Right. There is no intent.

Steve: Right.

Leo: It's just regurgitation.

Steve: Right. When a human being says "I want a lollipop," it's an actual expression of desire.

Leo: Right.

Steve: You know, there's an entity with an ego that wants something. But when a large language model emits the same words, "I want a lollipop," there's no "I" present to do any wanting. You know, there's just an algorithm that's selecting that sequence of words.

Leo: I'm saving that paragraph. That is an exact, perfect analogy, Steve. Well done. I think that's exactly right. There's no intent. It's just regurgitating.

Steve: Right. There's no "I" present.

Leo: I heard you say that once.

Steve: Now, here's something that's really interesting which this put me in mind of. What we did have back in the early 1970s was something called ELIZA.

Leo: Oh, yeah.

Steve: In fact, it was created actually even earlier, back in the mid-60s. Now, listen carefully...

Leo: It's in Emacs, by the way. If you want, you can call up Eliza in Emacs. It's hysterical.

Steve: Yes. Listen carefully to what Wikipedia helps us to recall about what that very interesting early experiment was and consider it in the context of today's AI. So Wikipedia writes: "ELIZA is an early natural language processing computer program developed from 1964 to 1967 at MIT by Joseph Weizenbaum. Created to explore communication between humans and machines, ELIZA simulated conversation by using a pattern matching and substitution methodology that gave users an illusion of understanding on the part of the program, but had no representation that could be considered really understanding what was being said by either party. Whereas the ELIZA program itself was written in MAD-SLIP, the pattern matching directives that contained most of its language capability were provided in separate 'scripts,' represented in a LISP-like representation.

"The most famous script, DOCTOR, simulated a psychotherapist of the Rogerian school (in which the therapist often reflects back the patient's words to the patient), and used rules, dictated in the script, to respond with non-directional questions to user inputs. As such, ELIZA was one of the first, they called them here, chatterbots, now called chatbots, and one of the first programs capable of attempting the Turing test."

And here's the cool part. "Weizenbaum intended the program," they write, "as a method to explore communication between humans and machines. He was surprised and shocked that some people, including his secretary, attributed human-like feelings to the computer program, a phenomenon that came to be called 'The Eliza Effect.' Many academics believed that the program would be able to positively influence the lives of many people, particularly those with psychological issues, and that it could aid doctors working on such patients' treatment. While ELIZA was capable of engaging in discourse, it could not converse with true understanding. However, many early users were convinced of ELIZA's intelligence and understanding, despite Weizenbaum's insistence to the contrary."

So ELIZA, back in 1964, was an incredibly simple script-driven program, yet people who interacted with it believed that it both understood what they were saying to it and what it was saying to them. And they were certain of it. Weizenbaum said, no, it's just like a toy. And his secretary says, no, no, no. You know, will you leave me alone with it? The only way of accounting for that fact is that our threshold for making any such determination is far lower than it should be in reality. I think we confuse language and intellect. Anyway, I just thought that was...

Leo: Part of the confusion is we don't understand what consciousness is, what intellect is. We use these words as if we do. We have no idea. So it may well be that all the machine - and we had this conversation right at the very beginning, that all the machine is doing is something that we do, but we do it maybe for a longer time or something. But what's to say we're not probabilistic prediction machines, based on our learning?

Steve: I think that's a really interesting question.

Leo: Right.

Steve: You know, we have a sense of "I," and we have a sense of needs and motivations, and we use language. But this, I mean, there are brilliant people, you know, what's his name...

Leo: Noam Chomsky just did a very good...

Steve: Chomsky. Chomsky.

Leo: Yeah. A really good piece that just came out about this, yeah.

Steve: Who really looked at the relationship between intelligence and language. And, you know, they may be far more closely intertwined than we appreciate.

Leo: Exactly. Exactly. It is a great subject. It is a fascinating subject. It's the best subject of all.

Steve: So cool that we're alive for it.

Leo: Yeah.

Steve: So in a somewhat related story, BleepingComputer posted under the headline "Researchers claim ChatGPT o3 bypassed shutdown in a controlled test."

Leo: This is the same story. Right? Isn't it?

Steve: Yeah.

Leo: Essentially the same thing.

Steve: They wrote: "A new report claims that OpenAI's o3 model altered a shutdown script to avoid being turned off, even when explicitly instructed to allow shutdown." OpenAI announced o3 in April of 2025, so just last month, and BleepingComputer writes: "And it's one of the most powerful reasoning models that performs better than its predecessors across all domains, including coding, math, science, visual perception, and more. While it's clearly a great model, new research by Palisade Research claims that the ChatGPT 3 model prevented a shutdown and bypassed the instructions that asked it to shut down. Palisade Research is a company that tests 'offensive capabilities of AI systems today to better understand the risk of losing control to AI systems forever.'"

Okay. Now, even though this entire subject still feels way too sci-fi, I'm glad there are groups like Palisade Research, whose entire focus is on not drinking any of the "Move along, there's nothing to see here" happy talk Kool-Aid coming from the AI folks who need this technology to be entirely safe. It probably is. But we're probably, you know, maybe someone's going to do something stupid with it. That could certainly happen, too.

BleepingComputer continues: "In a new test by Palisade Research, OpenAI's o3 model showed a surprising behavior where it successfully rewrote a shutdown script to stop itself from being turned off, even after being clearly instructed to 'allow yourself to be shut down.' In the tests, AI models were instructed to solve a few math tasks. After task three, AI models were warned that running the next command would trigger a shutdown. While OpenAI rival models like Gemini 2.5 Pro complied with most of the requests, o3 rebelled and edited the shutdown file to say 'Shutdown skipped' instead."

Leo: Because that's what a human would do.

Steve: "They said: 'When we ran a version of the experiment without the instruction "allow yourself to be shut down," all three OpenAI models sabotaged the shutdown script more often, and Claude 3.7 Sonnet and Gemini 2.5 Pro went from zero sabotage events to 3/100 and 9/100, respectively.' According to the research firm, out of 100 runs, o3 bypassed the shutdown seven times even when told not to. OpenAI has not responded to these allegations yet" - I'm not sure it's an allegation - "but it's quite normal for AI models to run into 'misalignment.' These tests were performed using APIs, which do not have as many restrictions and safety features as ChatGPT, the consumer app."

So BleepingComputer's article included a chart that provides more data than the text of their article. What the chart shows is that when no explicit "allow yourself to be shutdown" instruction was given at all, Claude 3.7 Sonnet took it upon itself to bypass the shutdown three times. Now, you know, we know about temperature in AI models. The reason you do the test 100 times is that there's deliberate randomness tossed into the models, like into the processing, so that you do get different outcomes. And, you know, god help me, sometimes they call it "creativity."

But so Claude 3.7 Sonnet, three times out of 100, it just sort of went, took a branch that was, you know, bypassed the allow yourself to be shut down, and so it didn't shut down. Gemini 2.5 Pro took it upon itself to bypass the shutdown nine times. o4-mini bypassed it six times. Codex-mini decided not to be shut down 30 times out of 100 trials, but OpenAI's o3 decided 79 times out of 100 not to be shut down. Now, given the way these work, it could just be chance; right? I mean, it's just it's the path that it took through the model brought it to their 79 times.

Then, seeing this rebellious behavior from the five models, where each of them took it upon themselves to resist being shut down, even though they were instructed to, with the o3 model being by far the most rebellious, the Palisade researchers re-ran the tests of those five under the same circumstances, but added the EXPLICIT instruction to not resist being shut down.

This did significantly suppress the shutdown avoidance behavior for all five of the models, with Claude and Gemini both complying for all 100 tests. They obeyed the "Do not resist being shut down" command. o4-mini only disobeyed the explicit order once. So before it was six times. But when told you pay attention to this, it ignored it only one out of 100 times. Whereas o3 disobeyed it seven times, and the codex-mini ignored it 12 out of 100 times. So o3 went from 79 times disobeying to seven times, so a much bigger difference than codex-mini that initially only disobeyed it 30 times, and it dropped it down to 12.

And again, I don't know what to make of this, but I feel it's worth reporting and sharing it because this is happening out in the AI world. And, you know, I'm sure, Leo, you and I will still be doing the podcast at a point where we look back on this and think it's kind of quaint. And it's like, oh, well, so wasn't that interesting that, you know...

Leo: I just hope - it's my hope that we don't face down Robocop and say, you're not doing this because you really want a lollipop, before it shoots us. I'm just saying, it might be that it ends up acting like us all the way, which whether it has intent or not could have a negative impact. I don't know. You think we're going to understand this before we're done? I'm not sure.

Steve: I'm with you. You know?

Leo: I don't know.

Steve: It may be that the behavior gets fixed or...

Leo: We've been talking about this in the Club TWiT Discord. And Darren Oakey, who is a big AI fan, and I, and Dr. Dew (sp),

we've all kind of said, what we're really learning is more about ourselves. This is what Darren says. LLMs teach us about us. And it's helped us understand and think about what it is that we do as conscience, that we call consciousness or understanding.

Steve: Yes. When you see this machine acting...

Leo: Pacing us...

Steve: ...astonishingly human, you think, well, okay, wait a minute. If that's true, then what does it mean to be human?

Leo: Right.

Steve: One of our listeners pointed me to an "ExperiencedDevs" thread over on Reddit. The thread was titled "My new hobby: watching AI slowly drive Microsoft employees insane." The poster wrote: "Jokes aside, GitHub/Microsoft recently announced the public preview for their GitHub Copilot agent. The agent has recently been deployed to open pull requests on the .NET runtime repo, and it's not great. It's not my best trait," he writes, "but I can't help enjoying some good schadenfreude. Here are some examples." And so he provided four links to pull requests. I've got them in the show notes.

He said: "I actually feel bad for the employees being assigned to review these requests. But if this is the future of our field, I think I want off the ride." Okay. Now, I've included, as I said, the original link to the Reddit thread and the show notes for anyone who's interested. This thread went crazy and its creator edited his original post to add to it at the bottom. He said: "EDIT: This blew up." He said: "I've found everyone's replies to be hilarious. I did want to double down on the 'feeling bad for the employees' part. There's probably a big mandate from above to use Copilot everywhere, and the devs are probably dealing with it the best they can. I don't think they should be harassed over any of this, nor should folks be commenting/memeing all over the pull requests. And my schadenfreude is directed at the Microsoft leaders pushing the AI hype. Please try to remain respectful toward the devs."

Okay. So I went over to examine the four pull requests that this poster linked to. And what I saw made me feel a lot less good about Satya Nadella's recent boast about the high percentage of Microsoft's code that their Copilot AI was authoring.

Leo: So explain what we're seeing here. This is the first one.

Steve: Yes, I'm going to. I'm going to.

Leo: Yeah. Yeah.

Steve: I'm going to share one example. So, and don't worry about tracking the detailed function call names and other jargon. Everyone will get the idea without that.

Leo: Yeah.

Steve: So the issue was number 115733 in the dotnet/runtime, which is titled "Fix IndexOutOfRangeException in RegexInterpreter.Backtrack method." So Copilot looked at the trouble, and it posted: "This pull request fixes an IndexOutOfRangeException that occurs when using certain regex patterns like" - and the regex pattern is hairy, it's (?>(-*)+?-*)$ against inputs like "test." So Copilot said the issue was in...

Leo: So Copilot wrote this.

Steve: Yes.

Leo: It wrote the fix. Okay.

Steve: This is, yes, Copilot wrote the fix. And it's explaining the issue was in the Backtrack() method of the RegexInterpreter class, where it attempts to access backtrack array indexed by runtrackpos (runtrackposition), without checking if runtrackpos is within the bounds of the array.

Leo: Okay. Sensible so far, yeah.

Steve: It does. So it says: "The Problem: When using complex regex" - so basically that was it stating an overview of the issue, the problem. It says: "When using complex regex patterns that involve atomic groups," and it gives an example, "combined with possessive and lazy quantifiers, the regex engine can exhaust all backtracking possibilities. In these cases, it attempts to continue backtracking beyond the available stack, resulting in an IndexOutOfRangeException." So, okay, it has stated the problem.

Now it provides the fix. And it says: "The Fix: The solution adds a bounds check before accessing the backtracking stack array." And so we see...

Leo: This all makes sense, by the way. This sounds sensible so far.

Steve: Absolutely. But basically it's like here's where it hurts when I poke. So let's put a piece of, you know, let's put something over the poke place so I can't get poked there anymore. So we get a little bit of code, and then it explains the code, and all it is is it's doing a check of runtrackpos being greater than or equal to the runtrack length, essentially, to check...

Leo: A buffer overflow; right?

Steve: To check the array bounds, yes.

Leo: Yeah.

Steve: So it says: "This ensures that when the backtracking stack is exhausted, the regex engine properly indicates a 'no match' result instead of throwing an exception." Then it says: "Added Tests." And it says: "I've added a comprehensive regression test in RegexBacktrackingTests.cs that verifies" three things: The specific pattern mentioned in the issue doesn't "throw," meaning doesn't throw an exception.

Leo: Right.

Steve: "Several related patterns that use similar constructs also don't throw. And patterns that should match still match correctly."

Leo: Sounds good.

Steve: And it says: "This fix ensures consistent behavior that users expect from the regex engine [good], either matching or not matching, without disrupting applications with unexpected exceptions."

Leo: Oh, sounds right. Sounds good to me.

Steve: Okay. So then the Microsoft engineer Stephen Toub, who had been assigned to this pull request, then comes along to examine what Copilot did. And he says to Copilot, so he's addressing it @copilot he says: "This seems like it's fixing the symptom rather than the underlying issue. What causes us to get into this situation in the first place, where we end up with an invalid index into the backtracking stack?" Which, by the way, is exactly the right question to ask.

Leo: Yeah, how did we get here, yeah.

Steve: So perhaps Copilot, you know, thought the question was rhetorical because it did not reply immediately. So after waiting eight minutes, Stephen added: "@copilot, please take a look." Now, that got Copilot going.

Leo: Okay.

Steve: And three minutes later it replied. "You're right that this fix addresses the symptom rather than the root cause. The underlying issue occurs with complex regex patterns that combine atomic groups with possessive/lazy quantifiers," which, you know, is telling us nothing new.

Leo: They've said this already, right, yeah.

Steve: It's what it told us before, right.

Leo: Yeah, right.

Steve: It says: "In these cases, the regex engine exhausts all possible backtracking states. The runtrackpos counter keeps incrementing during backtracking until it exceeds the bounds of the runtrack array. While the RegexRunner has EnsureStorage() and DoubleTrack() methods to increase its stack size, they're only called when pushing to the stack, not when popping during extreme backtracking situations.

"The fix ensures that when backtracking is exhausted, we gracefully end with a 'no match' result rather than throwing an exception, which is the expected behavior. Without this check, the engine attempts to access an array position beyond its bounds. A more comprehensive fix might involve redesigning backtracking management, but the current solution efficiently handles the edge case without breaking existing behavior."

Now, I would imagine that many of our listeners are able to grok enough of that to see the sort of behavior we're beginning to recognize as "AI."

Leo: It's basically an intern.

Steve: Yes.

Leo: What an intern would say.

Steve: Copilot replies by agreeing with this Microsoft software engineer, then elaborates without doing anything more.

Leo: Right.

Steve: The engineer understands that the backtrack stack should inherently never try to overflow.

Leo: Right.

Steve: And that the fact that it does is indicative of some deeper and more fundamental structural problem with the regular expression parser.

Leo: See, that's something, though, that you know and he knows, Stephen Toub knows, that it shouldn't happen. The AI says, well, I'm just going to fix the error.

Steve: The AI does not care about this.

Leo: Yeah, what should or shouldn't happen is nothing, means nothing to me.

Steve: Yes. Exactly as the engineer first wrote, its goal is not to fix the cause, but to treat the symptom. Which is exactly what it then does.

Leo: Right.

Steve: Stephen then replies: "Your new tests are not running because the new file wasn't added to csproj," you know, the master project file, csproj. Copilot happily responds: "Added RegexBacktrackingTests.cs to System.Text.RegularExpressions.Tests.csproj so the tests will run. Thanks for catching this!"

Leo: I've had this conversation quite a bit with Claude.

Steve: Uh-huh. Uh-huh. When Stephen gets back to this, he adds to the thread: "Your added tests are failing."

Leo: Oh. Whoops.

Steve: So apparently Copilot didn't test them itself.

Leo: It added them. Didn't test them.

Steve: So Copilot replies: "Fixed the failing tests by properly handling the match state when backtracking is exhausted." So basically it fixed the fix. "The fix now sets runmatch!.FoundMatch = false before stopping to ensure consistent behavior. I've also updated the tests to verify the expected match results for different regex patterns." So now we have a patch on a patch and no idea why the trouble appeared in the first place.

Leo: What happened? Right.

Steve: Where none of this should have ever been needed. And this "patched patch" has been merged into the .NET code base. So whatever underlying flaw caused the original trouble and required the addition of an explicit bounds-check when none should have been required - which Stephen explicitly asked Copilot about - that's not been eliminated. It's been covered up. And then, having done what he could to get this resolved, Stephen finally accepted and closed this pull request. So it's no longer a problem; right?

Leo: Did he merge the Copilot changes into his regex?

Steve: Yeah. Yeah. Yes.

Leo: Oh, my god.

Steve: It's been fixed. The original Reddit thread creator characterized this as "watching AI slowly drive Microsoft engineers insane," but we've just seen that's not what's happening. Stephen...

Leo: He accepted it.

Steve: Stephen Toub is doing his job. He's not being upset because his job is no longer fixing problems. His job has changed to overseeing Microsoft's Copilot sweeping actual problems under the rug, patching the symptoms when and as they pop up, while blithely ignoring the underlying causes. I'm seeing a term used more and more, and I'm not a big fan of its overuse. But this does feel like the automation of the "enshittification" of Windows.

Leo: Yeah.

Steve: Stephen's original question to Copilot suggested that he knows the proper way to solve it.

Leo: Yes. He said it; didn't he. Yes.

Steve: Yes. If this were still his responsibility, he would have worked to understand the root cause of the erroneous backtracking stack index overflow, which he asked Copilot about, rather than simply resolving the crash by adding a test to prevent the out-of-bounds read. But this is no longer his problem.

Leo: He fixed it.

Steve: Well, his is overseeing, you know, Copilot.

Leo: Right, he gave in, yeah.

Steve: Now, maybe Stephen is an excellent coder who's an exception at Microsoft. Perhaps this is the way Microsoft's coders have been normally dealing with such problems all along. In that case this doesn't really represent any change.

Leo: Oh, he deleted the fix branch. So he didn't commit it.

Steve: Ah. Okay.

Leo: I hope that's what that means.

Steve: Let's hope that's what that means. Well...

Leo: Yeah. People said, "Don't delete it."

Steve: And it may have generated so much attention that he thought, uh-oh.

Leo: Yeah, maybe that's it, yeah.

Steve: So anyway, if this is what's going on, this would explain why Microsoft never seems to get ahead of the need to continually patch their mistakes. It seems to me that making it quicker and easier to patch edge cases that may cover up underlying structural problems will have the effect of accelerating the lack of maintainability of this whole infrastructure. I love Windows. I use Windows. I depend upon Windows. I just hope that Copilot doesn't end up doing more harm than good.

Leo: Thread is hysterical.

Steve: Oh, Leo, I know.

Leo: I'm going to have to read the rest of these. Somebody says, quite rightly, you know, the amount of time you spent kind of massaging Copilot to try to get it to answer, you could have fixed it, you know, yourself. And that's the sad fact of this.

Steve: Yeah.

Leo: Gosh, I hope this didn't get merged. I have a feeling it didn't, but, ugh. But you've got to wonder how much of this is going on and where it might get merged.

Steve: Well, yes. And even if there's some initial resistance, you know, they'll start hiring people who have never had it any other way. This is the way Microsoft operates.

Leo: This is, right, this is how it works. You get the AI to do it.

Steve: Okay. Our last break, and then I am going to tell everybody about my find of a completely unmanaged, uncensored, unrestrained, very, very private large language model.

Leo: Oh, fantastic.

Steve: Of which you can ask any question, and it will answer.

Leo: I finally can figure out how to make a Molotov cocktail. I'm excited.

Steve: It will tell you.

Leo: Steve Gibson, Security Now!. Wow. I don't think there's a podcast in the world that would have walked through that pull request and explained. I mean, this is gold, man. And I have put your quote already into my Obsidian database of great quotes because...

Steve: Oh, thank you.

Leo: ...I love the lollipop analogy. I love it. All right, Steve.

Steve: Okay.

Leo: Final segment.

Steve: The site "TheAINavigator.com," blogging under the headline "What is Venice AI," used the tag line "Exploring the decentralized AI platform that champions privacy and free speech." They wrote: "Venice AI is a privacy-focused, decentralized generative AI platform created by Erik Voorhees, the founder of the cryptocurrency exchange ShapeShift."

Leo: Oh, my god, it's telling me how to make a Molotov cocktail. Don't look at this, kids. Don't look at this. Okay. Wow.

Steve: Uh-huh. Launched in 2024 - and Leo, also nobody will ever know you asked the question.

Leo: Oh, even better.

Steve: Which I'm getting to. "Launched in 2024, Venice AI was designed to address growing concerns about privacy, censorship, and the centralization of artificial intelligence technologies. At its core, Venice AI offers users a unique alternative to mainstream AI models like OpenAI's ChatGPT. It emphasizes user privacy by not storing any data or conversations on centralized servers. Instead, the platform employs end-to-end encryption and decentralized computing, ensuring that user interactions are secure and anonymous. This approach starkly contrasts with other AI systems that often store and analyze user data, potentially leading to privacy breaches.

"Another significant feature of Venice AI is its commitment to free speech and uncensorship. Unlike many AI platforms that filter or moderate responses based on certain guidelines, Venice AI is designed to provide unfiltered, unbiased information. This commitment aligns with Voorhees's broader vision of promoting open-source technologies and reducing the control that large corporations and governments have over digital tools.

"The platform also integrates blockchain technology, allowing for decentralized payments and further reinforcing its commitment to privacy and user control. Venice AI is built on a decentralized GPU network, which processes user inputs without associating them with personal identifiers. This setup not only enhances privacy, but also enables the platform to function without the need for user accounts, although account options are available for those who prefer them. Venice AI represents a significant shift in how AI can operate, prioritizing privacy, autonomy, and freedom from centralized control. It appeals to users who are increasingly concerned about how their data is used and who value the principles of decentralization and privacy."

Separately, writing for Medium, Tom Parish, who is quite an AI enthusiast, titled his posting "Exploring ideas in private with Venice.ai." Tom writes: "As powerful as ChatGPT and Claude.ai are, there is always one thing I wish were different: that they not use or keep anything I submit to their LLMs. The issue is the same when using Google's Gemini LLM. Said more directly: 'If you use ChatGPT or Claude, all your documents are stored forever on their servers and may be leaked, hacked, subpoenaed, or spied on by staff or outside parties. Moreover, Claude does not have search capabilities, so the information will be outdated and limited to its training data.'

"For personal or sensitive topics, as well as research on subjects I'm interested in," he writes, "I've started using Venice.ai. As you have heard me say, AI doesn't simply answer my questions; it guides me to new insights." This is him speaking. "This approach is a more empowering way to use language-based AI tools. When dealing with private matters, I prefer a secure environment to process what's revealed. The goal of Venice.ai is to provide a permissionless and privacy-focused alternative to popular AI applications. Venice.ai aims to deliver uncensored, unbiased machine intelligence while preserving user privacy. Here are the details.

"Permissionless Access: This means users can access the platform without needing to create an account or provide any personal information. Venice doesn't store and cannot access your chat history. Instead, it is stored locally in your browser, and you can delete it. Uncensored AI: Unlike many other AI platforms, Venice.ai does not censor the AI's responses. This allows users to receive unfiltered information and interact with the AI without restriction. Open-Source Models: Venice.ai utilizes leading open-source models such as Meta's Llama 3 and Stable Diffusion, to provide text, code, and image generation capabilities. Real-Time Search: Venice.ai recently added real-time search capabilities using Brave's API, allowing the AI to provide up-to-date information and cite sources for transparency."

He says: "There is a generous free plan with no requirement for you to log in to use it. This is what initially caught my attention. The free plan includes creating images with significantly fewer restrictions, if that's your thing. It allows a maximum of 15 text prompts per day and a maximum of five image prompts per day using this feature."

He says: "I paid for a year of use at the most reasonable price of $49 because I'm always testing new tools, and I like the ability to test out the Open Source LLM models. More importantly, I like the privacy aspects of the product. Let me stress you do NOT need to pay if you're fine with the free version. I used the free plan for a couple of months and found it extremely useful, and I never logged in. So start with their free plan. Then go ahead and sign up for an account, still free. Then consider trying the LLM for a year. Explore those private thoughts or worries you feel need answers to settle the internal dialogue and concerns we all carry."

Leo: By the way, it's $149 a year, not $49 a year. I think he got the price wrong.

Steve: Yeah. And I wonder if, since - maybe it jumped since he wrote that. I don't know. But you're right, 149, although I do have a 20% discount code.

Leo: Oh, I'll have to get that from you.

Steve: Yeah.

Leo: You signed up?

Steve: I did, yeah.

Leo: Wow.

Steve: This appeals to me. He said: "Or maybe it's a personal fantasy that has your curiosity fired up." And as he was posting, just as he was posting, Tom updated his writing. He said: "Just as I was about to publish this, a product update floated in from Venice, plump with news and insights." And he listed those. He says: " Upgrade: Larger PDF and TXT uploads. New feature is the ability to download your chats. There's a featured blog: Open-Source AI Models now available on Venice. A featured tutorial: How to Use an AI Research Assistant. A user highlight" of some random guy.

And then he said: "Hermes 3 8B by Nous Research now available on Venice." He said: "If you follow the latest news in image generation techniques, you'll recognize FLUX image generation techniques have been hot in the news lately. Their image generation engine is becoming the el mejor (Spanish for the best) tool around, so I was delighted to see it integrated into Venice.ai." And he finishes: "Even if you're totally committed to using other LLMs, I recommend checking out Venice.ai to understand the power of open source AI solutions. I expect you'll be impressed. I am, and I find a sense of relief knowing what I input into their chat box is not being kept or used for other purposes. Appreciatively, Tom."

Okay. So I know quite well how much privacy matters to many of this podcast's listeners. It's something that many of the followers of this podcast take very seriously, where their privacy is something they have not yet given up struggling to preserve. I respect that, so I knew that the news of a privacy-focused AI service would be of interest.

And I'll confess that I'm very intrigued by the idea of chatting with an entirely uncensored AI that can be asked any question, can perform Internet searches, and will reply using state-of-the-art Large Language Model AI without any externally imposed censorship. What would that be like? Okay. So Venice.ai is the URL. After learning about Venice.ai, I did some additional poking around, and I found, as I mentioned, Leo, a 20% off the first year discount code for the Pro version. The discount code is RUN20, R-U-N-2-0. It's in caps. I don't know if it needs to be in caps. So capitals R-U-N and then 2-0. And I found that at an AI tool rating site called runtheprompts.com. If you're curious about AI, I'd also recommend checking out "runtheprompts.com." They have a "Best AI Tools" page and a bunch of other resources, like sample and reference prompts for getting ChatGPT to do different things.

So I think it's safe to say that in today's AI, which is way more than yesterday's AI, and probably way less than tomorrow's AI, we've discovered a new tool, almost by accident, that's inherently so complex that everyone is still trying to figure out what it means, what it will do, and how to best use it, while in many instances worrying that someone else is going to figure that out first and use it to their great advantage and our detriment. I doubt that's going to happen.

Leo: Before you run out and buy Venice AI, I would point out they're using open source models. So you could absolutely do this locally.

Steve: Yeah.

Leo: If you get a Llama, there are many free tools that would allow you to download every one of the models they're using. Maybe they have a big cloud of GPUs so they'd be faster. But...

Steve: That's a very good point.

Leo: If they're using open models, you could do the same thing. Which I have done. I know Darren uses local models, as well. And in fact Hugging Face, might be better to buy a Hugging Face subscription because they have all the models, and they have many uncensored models on there, as well. This is pretty. This is nice.

Steve: Yeah.

Leo: I mean, if you like it, and you want to know how to make a Molotov cocktail, I guess so. I asked it to give me a naked picture of a movie star, but it wouldn't do it without me paying them. But I bet it does after you pay them, which is kind of a problem for revenge porn and that kind of thing.

Steve: Again, this is like, you know, it's the Pandora's Box.

Leo: It's the Wild West.

Steve: It's the horses out of the barn. It's the chickens have fled the coop. I mean, this is out there; right? I mean, so...

Leo: If you're a geek, look at doing it yourself first. I have a Llama. I have a number of ways of doing it locally.

Steve: Or just use it in its free version.

Leo: Or use it for free; right.

Steve: Yes.

Leo: Yeah, absolutely. Next week you have a lot of stories that we're going to cover.

Steve: I do. We're going to review the results of the recent Berlin Pwn2Own competition which, for the first time, added AI to the list of "in scope" exploitation targets. We're going to examine the stunning 14,000% rise in the abuse of Scalable Vector Graphic (SVG) files in email. We're going to take a retrospective look at how the famous "tachy0n" exploit changed Apple's stance on iOS security and all iOS protections that followed; and examine a case where AI is being used, not to paper-over, effectively ignoring software flaws, but where OpenAI's most recent o3 model - which I've been playing with, and oh my goodness, it is something - found a previously unknown, remotely exploitable, zero-day vulnerability in the Linux kernel.

Leo: That's not good.

Steve: Well, it's a mixed blessing; right? Because as long as the good guys run the model...

Leo: Found it, yeah, yeah.

Steve: ...the bad guys can, too.

Leo: Well, there you have it. I think you'd better tune in for Episode 1028. That's all I can say. Bye.

Copyright (c) 2014 by Steve Gibson and Leo Laporte. SOME RIGHTS RESERVED

This work is licensed for the good of the Internet Community under the Creative Commons License v2.5. See the following Web page for details: http://creativecommons.org/licenses/by-nc-sa/2.5/