



Daybreak & Codename MDASH

Description: Microsoft rethinks Edge's "intended behavior" after it gets press. Chaotic Eclipse hacker strikes again with a BitLocker bypass. Google's threat analysis group documents malicious AI use. Canada hasn't learned the lessons of the EU and the UK. AI chatbots may be far more addictive than social media. "Project: Hail Mary" now available to stream. An apparently serious zero-point quantum vacuum energy source. A bit of listener feedback. OpenAI's and Microsoft's vulnerability discovery systems.

High quality (64 kbps) mp3 audio file URL: <http://media.GRC.com/sn/SN-1079.mp3>

Quarter size (16 kbps) mp3 audio file URL: <http://media.GRC.com/sn/sn-1079-lq.mp3>

SHOW TEASE: It's time for Security Now!. Steve Gibson is here. A little change in Microsoft Land over that Edge password thing. We will talk about a new way of making chips work without electricity by pulling in quantum power from the air? Is that possible? And OpenAI and Microsoft's response to the Anthropic Mythos security tool. All that and a whole lot more coming up next on Security Now!.

Leo Laporte: This is Security Now! with Steve Gibson, Episode 1079, recorded Tuesday, May 19th, 2026: "Daybreak and Codename MDASH."

It's time for Security Now! - yay! - the show where we cover the latest security, privacy, computer, sci-fi, everything on this man's mind. Mr. Steve Gibson is here. Hello, Steve.

Steve Gibson: You know, Leo, we were just using the expression "has a mind of its own." And I realized, we really can't say that any longer without meaning it because things do or very soon will actually have a mind of their own.

Leo: Have a mind of their own? Yeah. Like your car.

Steve: Yeah, I mean, cars are - yeah, exactly.

Leo: That's a really interesting - you know, this is one of the big debates that's going on is, is AI conscious? And in fact it's one of the first questions I asked you when we started talking about AI on this show is where you stood on that. And you, correct me if I'm wrong, but I think your position is the same as mine, which is there isn't anything special going on inside our brain that couldn't be duplicated by a physical process outside of our brain. It may not be yet, but...

Steve: I believe that's - yes. I was talking to somebody who's not a techie yesterday, who was interested in the topic. And the way I framed it, I think, I know it worked for him. I said, "AI is language. And language is knowledge, but not understanding." And when he kind of looked at me, I said, "Think about a book. A book is language printed on paper. So obviously a book contains knowledge. A book has knowledge."

Leo: No understanding.

Steve: But, exactly, no understanding.

Leo: Right.

Steve: And I said - because I've been in computing my entire life. So what I'm interacting with, in this case Claude, I'm still like stunned by it. In fact, I have a little one-pager editorial about my feelings after the last week of the danger that we are in, not the kind of, well, maybe some people are worrying about it. But how seductive and addictive it is. It is inherently that. And if we thought social media was a problem, baby, you ain't seen nothin.'

So, but anyway, so I said, you know, in watching AI, I can see, when I see its mistakes, I realize that it reveals it doesn't understand what it's producing. It's producing astonishing content, but it doesn't understand it. And so that, when that changes - and I agree with you, Leo, I don't see any reason why it can't. I don't know when or how or what. But, you know, and this whole LLM era may just be, you know, the beginning of this. Lord knows, you know, anybody, cancer researchers, fusion researchers, quantum computing researchers, they all say just give us money.

Leo: We'll figure it out.

Steve: And we can make it happen. Well, we've never seen anybody give anything money more than AI. I mean, this is just ridiculous. So if there's an answer, and if money can find it, then we're going to have an answer. I mean, we're going to see this thing continue to go.

Leo: Yeah.

Steve: Because - and I have to agree, you know, if you - I think you were referring to it, that last scene toward the end of the "Wall-E" movie. I haven't seen it for a long time. But it was a bunch of obese adults floating on a Starliner.

Leo: Sucking on their smoothies.

Steve: Like they were so fat that their bones were like being pulled apart or something. I don't quite remember what the visual was. But, you know, and "The Matrix," right, everybody in a pod who doesn't know that they're not...

Leo: They're just batteries.

Steve: Yeah. And so imagine if, well, anyway...

Leo: We may be headed there, is what you're implying.

Steve: Something is - this is a problem for us. Anyway, so not surprisingly, today's topic is Daybreak and Codename MDASH.

Leo: Oh, boy.

Steve: Yup, there it is.

Leo: There they are floating down the...

Steve: Oh, goodness, yeah, just, yeah.

Leo: It was a wonderful movie, actually.

Steve: Yeah, it is good. So Daybreak and Codename MDASH, which are, you know, the responses to Mythos in various ways. Also we're going to talk about how - so we'll get to that at the end. But first Microsoft has decided to rethink Edge's so-called "intended behavior" after it got some press.

Leo: We didn't intend that intended behavior.

Steve: Not favorable, yes.

Leo: After all.

Steve: Speaking of Microsoft, the Chaotic Eclipse hacker has struck again with a bypass of BitLocker, which some people have called a "backdoor." I think that's taking it too far. Also Google's Threat Analysis Group documents their discovery of the clear malicious use of AI, which we're beginning to see. Apparently Canada has not learned the lessons of the EU and the UK, so their Parliament is going to go down that same rabbit hole of, you know, legal disclosure and tapping and so forth. We'll talk about that. I want to take, as I said, a moment to talk about how AI chatbots may be far more addictive than social media and why I think that is probably going to happen. Also, a comment about a favorite piece of sci-fi of ours, "Project: Hail Mary," now being available to stream.

Also I put this out there just because it was fun, and it is so wacky and interesting, an apparently serious zero-point quantum vacuum energy source. And every so often I hit a nerve among our listeners. And, boy, you know, thanks to the fact that these notes went out early on Sunday, there's been a lot of time for some feedback from our listeners. So

we're going to have fun with that, and actually share some feedback, and then take a look at OpenAI's and Microsoft's vulnerability discovery systems.

Leo: Oh, good. Oh, good. Yeah, I mean, it was pretty clear, we talked about this just a couple weeks ago, that Mythos is very effective. There was just a story last week about discovering a flaw in macOS which is pretty darn locked down, getting around Gatekeeper. So, yeah.

Steve: Yeah.

Leo: There's definitely some stuff. Oh, and by the way, Steve, there is a picture of our future here in the Club TWiT Discord.

Steve: Uh-oh.

Leo: I'll pull this up for you, and you can see it. I think this looks good. I think this is maybe our retirement plan or something like that. I don't know. I'm just saying. Oops. Squish you down so there's room for us in our hover chairs. That's a podcaster's dream right there.

Steve: I like it.

Leo: But I need that smoothie. Get on that right away.

Steve: And I'm sure you've noticed there's been a complete revolution in this sort of thing, like ads now look different.

Leo: Oh, yeah, everything.

Steve: Like late-night comedy sketches are now using entirely different imagery because it's - now you...

Leo: So easy.

Steve: ...don't need to have a huge staff of artists in order to create something.

Leo: Darren, what did you use? Because this is - is this Nano Banana? Because this is really - looks really good, I have to say. Oh, he says ChatGPT, interesting. They're all doing it now. Google's doing it now, too, with, you know, they're doing agents. I mean, it's amazing.

Steve: Yeah.

Leo: Anyway, let's take our first break, just to get this out of the way so we can get to the Picture, the long-awaited Picture of the Week, in just a little bit. I haven't seen it. I closed my eyes. But we'll see it together in just a little bit. Picture of the Week time, Steve.

Steve: So in keeping with today's podcast theme, I gave this picture the caption "Worries over AI surpassing us may be overblown because AI has been trained on human output."

Leo: Uh-oh. Let's scroll up here. Oh, this is wrong on so many levels. So many levels.

Steve: So this, I can't explain this. But then again, if AI is trained on us, then I don't think we have anything to worry about. We see the right side of a gate which is open at the moment. The sign very clearly states "Please close the gate to keep the seagulls out." Now, you know, last I checked, seagulls could fly. If it said "the chickens," you had to keep the chickens out, okay, you know, a flightless bird, that would make sense. Here it's not clear how having a gate closed would affect seagulls one way or the other. I mean, unless they like to walk. Anyway, yes, I don't know what's going on here, Leo. But if AI is at our level, being trained on our output, then I think we're going to be fine for a while.

Leo: Not to mention the fact that the gate doesn't go all the way across the gap, either.

Steve: I don't know what the hell.

Leo: The whole thing is just screwy.

Steve: Yeah.

Leo: That's really funny. I love it. All right.

Steve: So last week, we noted the discovery, the reporting, and the widespread confirmation among some of our own listeners that Microsoft's Edge browser, remember, was storing all of its users' passwords in RAM, in plaintext, decrypted, just sitting there, where they were easily discoverable and exfiltratable en masse. The data included the URLs so you knew where to go, the usernames and passwords so you knew what to put in once you got there, which were required to log into every website whose data was present in Edge's password list, and presumably where no other authentication factor would be required.

Now, this brings me to something we've talked about before. I'll just, you know, take a little segue here to pause and note that this is a perfect example, that is, Edge doing this, having this heinous behavior, a perfect example of the reason why, if one is going to go to the trouble of having additional factors of authentication security, it's nuts to store that additional authentication information with the same single provider as the other, as your other authentication information is stored. Our listeners have asked, you know,

several times whether it's safe to store their one-time password secrets in the same password manager as their usernames and passwords.

You know, this comes down to the meaning of the word "safe." They want me to say yes because it's so convenient to extend a password manager's capabilities to include responding to the query for a six-digit one-time-password token. I really do get it, and I understand the temptation here. So I'll just say that I've never done that, and I never would. The entire point here is separation and redundancy, which is completely lost when all of the eggs are stored in a single proverbial basket. You know, I use, as I've said, OTP Auth, nice little iPhone app, iOS app on my - and iPad - on my separate iPhone.

The good news is that most sites have become much smarter about avoiding needless prompting for one-time password tokens. Whereas, you know, a financial institution or the government might reasonably insist upon the provision of a one-time password every single time you log in, or maybe if you haven't touched the site for even 30 minutes or so, you know, many other less sensitive sites that have been configured to require a one-time password will nevertheless relax their need when the browser being used already carries a previously valid login cookie, which indicates that that browser was previously logged into that site.

You know, this is the newer "we recognize you on this computer" messaging that we're seeing more and more often now. And that's good, right, since we want bad guys who will not have that browser cookie to be forced to come up with that additional authentication factor, whereas we don't want it to be overly burdensome for regular users who want that added safety without the overboard hassle.

Anyway, my point is, here's an example. You know, if one-time password secrets were also exposed by Edge, as presumably they would be if Edge were to support that, then it would have been the keys to the kingdom. If, however, somebody had kept their one-time passwords anywhere else, then they would have still had protection for all the sites that they cared enough about to establish a one-time password. So, you know, and again, if you really don't want the security, go ahead, store them all in one place, and you get the convenience of a password manager that does all that for you. But not me.

Okay. So getting back to Microsoft and Edge. Last week we noted that Microsoft's disappointing but predictable response to questioning about their in-the-clear storage of the users' authentication data was that it was "intended behavior." Yes, that's what we intended. We intended it to be all out there in RAM so anybody could get it. The SANS, remember, the SANS Security Institute wrote: "Microsoft classifies this as 'intended behavior.'" And the guy writing for SANS said, "I'm not sure what manager or lawyer decided that. Hopefully it wasn't anyone in their security team." Amen. Since I titled this first bit of news "Intended behavior only until it gets media attention," you can guess what comes next; right?

BleepingComputer provides the details and the background, writing last Friday: "Microsoft is updating the Edge web browser to ensure it no longer loads saved passwords into process memory in clear text at startup after previously stating it was 'by design.' This behavior was disclosed on May 4th by a security researcher Tom Ronning, who demonstrated that all credentials stored in the Edge built-in password manager were decrypted on launch and kept in memory, even when not being used. Ronning also released a proof-of-concept tool that would allow attackers with admin privileges to dump passwords from other users' Edge processes. Those without admin privileges would only be able to dump them from their own. He said he reported the issue to Microsoft and was told the behavior was 'by design' before he publicly disclosed it."

And I'll note that this is an interesting wrinkle on the "responsible disclosure" principle; right? You tell someone responsible, like Microsoft, in confidence, about some clearly bad

behavior you've just discovered in one of their highly security critical flagship products, and you're quite clearly told, "Yeah, that's right. That's what we want, so that's the way it is." Okay. At that point, no one is going to fault you for letting the rest of the world know what you have found and that you were basically told to buzz off.

BleepingComputer quotes the discoverer, saying: "Edge is the only Chromium-based browser I've tested that behaves this way. By contrast, Chrome uses a design that makes it far harder for attackers to extract saved passwords by simply reading process memory."

Bleeping Computer wrote: "While it initially refused to address the issue, telling BleepingComputer at the time that 'this is an expected feature'" - that's right, it's not a bug, it's a feature - "'this is an expected feature of the application,'" they said: "Microsoft announced on Wednesday" - so that's, you know, six days ago - "that future versions of Edge will no longer load saved passwords into memory on startup, even though the reported scenario falls within the expected existing threat model, which excludes attacks where an adversary already has administrative control of a device."

They wrote: "Microsoft Edge Security Lead Gareth Evans said: 'This defense-in-depth change'" - meaning what they're going to change Edge to do, certainly not what it had been doing, which they were previously defending - "'now this defense-in-depth change will come to every supported version of Edge (Stable, Beta, Dev, Canary, and the Extended Stable channel our enterprise customers run),' he said, 'and we're prioritizing the rollout.'" All right. Now that everybody knows and is upset and is writing in about this, they're going to change it posthaste.

"'With our commitment to the Secure Future Initiative and customer feedback, we are taking a broader view. Well. That means looking not only at whether something meets the bar for a security issue, but also at where we can reduce exposure through defense-in-depth improvements. In this case, reducing the exposure of passwords in memory is a practical step in that direction.'" It's almost as if, Leo, nobody thought about this before. They're just like, eh, what, you know. And then when someone said, what about that, they go, ooh. Yeah, yeah, we should probably change that.

Leo: Oh, you want defense in depth. Oh.

Steve: Yeah, in depth is, oh, yes. We thought you meant death. No, not death, depth.

But anyway, they said - Bleeping Computer wrote: "The fix is already live in the Edge Canary channel and will be included in the next update for all supported Edge releases (from build 148 and newer)." They said: "Last year, Microsoft introduced a new Edge security feature to protect users against malicious extensions sideloaded into the web browser, and restricted access to Edge's Internet Explorer mode after hackers began leveraging zero-day exploits in the Chakra JavaScript engine to access target devices."

Okay. So first, while writing this on Saturday, I immediately fired up Edge to check its Help/About, and I watched it quickly updating itself to build 148. So that fix was, indeed, quickly pushed out. Everybody has it now. Or if you haven't run Edge for a while, you will immediately upon launching it the next time.

But the point Microsoft made about the threat model governing Edge's design was important. I think it's reasonable, and it's worthy of a little bit of additional attention. BleepingComputer, remember, wrote: "Microsoft announced on Wednesday that future versions of Edge will no longer load saved passwords into memory on startup, even though the reported scenario falls within the expected existing threat model, which excludes attacks where an adversary already has administrative control of a device." In

other words, they're saying - I'm making this up - we're going to change this behavior even though the scenario Tom Ronning discovered, where all username and password authentication was being needlessly pre-loaded into RAM, does fall within the expected existing threat model.

Okay, now, first, before I defend Microsoft's response, I'll take exception to their use of the term "administrative control of a device." As was noted, administrative control is explicitly not required. Administrative control allows malware to obtain the usernames and passwords, or I should say allows malware to also obtain the usernames and passwords of ALL of a system's users who may be logged in at the time in other sessions that has Edge running. But malware running in a non-admin account can still access all of its own users' in-RAM Edge authentication. So, eh, not quite right there.

But let's focus upon the intent behind Microsoft's defensive position. The concept and deliberate design of formal threat models is perhaps the most important advance in our understanding and practice of security. We saw a lot of that during last week's deep dive into DigiCert's internal security architecture. Just the fact that an "architecture," the word "architecture" is something that security can now have, that represents a significant advance in our state-of-the-art understanding of how to provide protection. You know, a lot more theoretical thought and modeling has gone into modern security understanding, the fact that we have, you know, the notion of, as I said, an architecture.

So in this case Microsoft is essentially saying: "We recognize that once an attacker has taken up residence in a system, by whatever means, our ability to limit the damage that could be done is severely limited by the tradeoffs we have had to make in the name of practical usability." What comes to mind immediately is User Account Control. I may refuse to store my one-time password secrets in my password manager, just as a matter of principle, but the first thing I do when setting up a new Windows machine, before I totally lose my mind, is completely disable UAC. Having that thing constantly darkening my doorway - I mean, my screen - and popping up to get my permission when I want to do perfectly safe things, the consequences of which I perfectly understand, is not offering any value proposition that works for me.

I get it that for the typical Windows user, yes, you need to have a nanny looking over your shoulder all the time. But, you know, no thanks. My sanity is important to me. So UAC? I'll take responsibility for turning that off because I want to get work done. And as a developer I'm doing a lot of things that your typical Windows user doesn't. But I am appreciative of the fact that Microsoft is in an impossible position, that is, trying to secure people who are going to fight against that. So to that end I am sympathetic. Windows is being used by people who will follow commands provided to them by some random page on the Internet, instructing them to blindly paste and run a command they could not possibly understand, even if they could see it. So how is Microsoft supposed to protect such users from themselves when an increasingly hostile world wants to attack them?

So on the one hand, Microsoft's position that there can be no true protection from bad guys who have already gotten into one's PC, you know, it's accurate, and it's defensible. In fact, in a minute or two we're going to examine what's been dubbed "The BitLocker Bypass." It's a perfect case in point about the nature of local compromises and security boundaries. And a security boundary is another new theoretical concept that we didn't have, you know, originally, which is part of modern security architecture.

But the other point Microsoft made, quoting the phrase "defense in depth," refers to another of the crucial advances that have been made in our contemporary understanding of security. When a castle was surrounded by a piranha-filled moat, attackers could just bring a boat and float it across the moat. But when the outside of that moat is surrounded by a tall fence, then it would be difficult to get the boat to the moat. So "defense in depth" is also exactly storing all authentication factors in a separate location

because storing them in the same place is sacrificing the opportunity to have additional depth.

So in this case the bottom line is that the attention drawn to Edge's entirely needless exposure of its usernames and passwords - and notice how quickly they fixed it. I mean, it's not like this took a couple months to get right. I mean, it's like, oops, and like the next day they had an update ready, and they pushed it out to everybody without any testing needed because it was simple to do. They just hadn't. So that exposure was needless. As we saw, none of the other Chromium-based browsers ever behaved so cavalierly with their users' most important secrets. So every one of those took the time and trouble to protect them. Now Edge does, too. So that's good.

And Leo, you know what else is good? I need to take a sip of coffee, and we're a half-hour in.

Leo: Fair enough. Fair enough.

Steve: Let's take a break, and then we're going to talk about the recently discovered bypass of BitLocker's encryption. Was it a directly planted backdoor, or what?

Leo: Yeah, because some people have said that. Well, that's a backdoor.

Steve: Yes. That's what it's called.

Leo: You know, because I have kind of - Google said this, too. If somebody's in your computer, whether the passwords are in the clear or encrypted, they're in your computer, you're in deep trouble.

Steve: Yes. And that is a good point.

Leo: Isn't that the antithesis of Zero Trust? I mean, Zero Trust says if somebody's in your network, it doesn't mean that they should have free rein now. You can't, you know, you still want to put some - it's layered security. You still want to put some barriers up.

Steve: You know, our topic at Zero Trust World, right, the call is coming from inside the house means even if you've got a bad guy in your home, you have segmentation so that, you know, you have put up barriers inside that prevent them from going where they shouldn't.

Leo: Limit what they can do.

Steve: Yeah.

Leo: And that seems pretty reasonable.

Steve: And the problem is the tradeoff for convenience. We're always hitting that wall. We're always saying, I mean, you know, we've talked about it. It's kind of cool to put in your magic six-digit code. You're like Bond; right? You know, it's like, oh, what's my code, in order to get authenticated. I mean, it feels more secure. And in this case it is. But you shouldn't have to do it every time you look around.

Leo: It's funny, you turn off UAC, I was thinking about that's how I use AI. I use what they call YOLO mode, which is I say, yeah, do whatever you want. I don't have time to approve very darn bash command. Just go ahead. I trust you. What could possibly go wrong? Back to you, Steve.

Steve: Okay. So while we're on the topic of Microsoft, and we'll get back to it at the end because MDASH is their vulnerability, their internal proprietary vulnerability finding AI system. But for now I want to make sure that everyone knew about the recent discovery, with a published proof of concept, of a local bypass attack on Microsoft's proprietary BitLocker drive encryption.

The source and the apparently deliberate timing of the disclosure of this latest significant Windows vulnerability is interesting because it was publicly released last week on the 13th, the day after this month's Patch Tuesday. So Microsoft couldn't fix it for the previous day. And who released it? None other than the hacker Chaotic Eclipse with his Nightmare-Eclipse GitHub account. Remember that this is the individual we talked about recently who is extremely perturbed by Microsoft's...

Leo: Oh, this guy.

Steve: Yeah. Extremely perturbed by Microsoft's handling of him and his disclosures. Recall that he appears to accuse and blame Microsoft for deliberately and knowingly ruining his life. I mean, like, words to that effect.

Leo: Wow.

Steve: I mean, he's like, what? And he's never really exactly clear what it was. But it's like he was counting on the reward which he says they deliberately denied him, and so he wasn't able to meet other commitments that he had already, like, pre-banked. Who knows? But anyway, in retaliation for that perceived slight, he has previously disclosed the BlueHammer and the RedSun local privilege escalation vulnerabilities as zero-day flaws, saying ta-da, here you go, with proof of concepts, and they were immediately exploited in the wild shortly after he disclosed them.

So now, same guy, Chaotic Eclipse is back, publishing two new exploits with proofs for two new unpatched vulnerabilities named YellowKey and GreenPlasma. They are, respectively, the BitLocker bypass; and the second one, GreenPlasma, is a privilege escalation. He describes the BitLocker bypass issue as functioning like a backdoor because the vulnerable component is present only in the Windows Recovery Environment (WinRE), which is used sort of as a utility host OS. It's that reserve partition that Windows now establishes when you're installing Windows onto an empty hard drive, that allows you to boot into some special place. It's often used to repair boot-related problems with Windows. When the rest of the OS won't boot, you're able to use this recovery environment.

So this Chaotic Eclipse guy remains miffed at Microsoft and has published guidance on how to exploit this hole that he has found. And if that wasn't enough, he has promised what he described as "a big surprise" for the next Patch Tuesday. So a couple weeks from now we may get something else.

The security researcher Kevin Beaumont, who posts as "GossiTheDog," has independently confirmed the functioning of the YellowKey BitLocker bypass. Kevin's first post over on Mastodon was: "So I've just had a quick play with this; and, yes, it works. Essentially BitLocker" - this is Kevin Beaumont saying this. "Essentially BitLocker has a backdoor. Mitigation," he says, "equals BitLocker PIN and BIOS password lock." Okay, now, of course, a BIOS password lock is a pain in the butt because you've got to enter it every time you turn the computer on. But for high-risk scenarios where local access with rebooting might be possible, that is, where someone could get a hold of a computer and reboot it, because that's what this requires in order to get access to BitLocker, the BIOS password lock would be the strongest and the quickest cure until Microsoft arranges a fix for this.

Kevin followed his first Mastodon posting with a thread of posts which I've collapsed to read, he wrote: "I think my prior toot on Nightmare-Eclipse auto deleted. So to make a perm one," he said, "I suspect it's somebody who used to work at Microsoft, who departed after my era. For anybody looking at this, testing showed two things: TPM unlocked the storage. It provides a login bypass, as you're dumped as SYSTEM prior to Windows Hello or password login." He says: "BitLocker operates without a PIN by default, so it's basically a big gap. It's unclear how this code made it into the production version of Windows. I should point out I've only tested with one version of Windows 11. Maybe the scope is smaller. Will Dormann and I have both recreated the BitLocker backdoor, er, vulnerability."

Okay. So what's the story? BleepingComputer's headline, and that's where Will Dormann comes in, was "Windows BitLocker zero-day gives access to protected drives, proof of concept released." Since we already have a lot of background, I'm going to skip over their description of the trouble and excerpt just some of the good bits. They write: "The researcher says that YellowKey is a BitLocker bypass that affects Windows 11 and Windows Server 2022 and 2025. It involves placing specially crafted 'FsTx' files on a USB drive or EFI partition, rebooting into WinRE, and triggering a shell by holding down the CTRL key. The BitLocker bypass should also work without USB storage by copying those files to the EFI partition on the target drive. According to Chaotic Eclipse, the spawned shell gains unrestricted access to the storage volume protected by BitLocker." In other words, when you do this, the volume is not encrypted. It's just there.

So they write: "Independent security researcher Kevin Beaumont confirmed that the YellowKey exploit is valid and agreed that BitLocker has a backdoor." Okay, we'll talk about that in a second. They write: "He recommended using a BitLocker PIN and a BIOS password as a mitigation. In an update, Chaotic Eclipse said that 'the real root cause is still not known by the general public,'" and then BleepingComputer continues, "and that the vulnerability is exploitable even in a TPM - Trusted Platform Module - and PIN environment."

They write: "However, the exploit for this version has not been released. The researcher said: 'I think it will take a while even for MSRC [Microsoft Security Research] to find the real root cause of the issue.'" I don't think so, but that's what he said. He says: "I never managed to understand why this vulnerability is so well hidden."

Okay, so note that the term again, "backdoor," keeps floating around this, which I would call a "vulnerability." Kevin carefully noted that "it's unclear how this code made it into the production version of Windows." And if Chaotic Eclipse is correct, which I'm suspicious of, that there's also a full PIN protection bypass - and again, I suspect that's a

specious claim - then it would make for a powerful backdoor for BitLocker. But that's a lot of ifs. BleepingComputer reports Chaotic Eclipse saying: "No, TPM+PIN does not help. The issue is still exploitable regardless. I've asked myself this question, can it still work in a TPM+PIN environment? Yes, it does. I'm just not publishing the proof of concept. I think what's out there is already bad enough."

Okay. Maybe. But to me it feels out of character for Chaotic Eclipse, given everything we know about this individual, to willingly hold anything back. What's the point? Once Microsoft fixes the vulnerability, the problem, with or without the PIN, will be resolved. So it's not as if holding onto another aspect of the bypass would have any future value. In any event, I mean, again, I think Chaotic Eclipse is boasting and bragging beyond what he actually has.

In any event, BleepingComputer continues, saying, "Will Dormann, principal vulnerability analyst at Tharros Labs, also confirmed that the YellowKey exploit worked with the FsTx files on a USB drive, but could not reproduce the bug using the EFI partition. He explained to BleepingComputer that: 'YellowKey exploits NTFS transactions in combination with the Windows Recovery image. This PIN prompt happens before Windows Recovery is entered.' Dormann clarified the exploit process, saying that to boot Windows Recovery, 'Windows looks for \System Volume Information\FsTx directories on attached drives, and will replay any NTFS logs.' The result of this is that the X:\Windows\System32\winpeshl.ini is deleted; and when Windows Recovery is entered, rather than launching the actual Windows Recovery environment, it pops up a CMD.EXE with the disk still unlocked."

They said: "By default, TPM-only BitLocker configurations, meaning those without a separate PIN, unlock encrypted drives automatically without requiring user interaction." Now, what they mean is just like in the normal course of events, you come into your office in the morning, you turn on your computer, that's what happens. TPM-only BitLocker configurations, meaning those without a separate PIN, unlock encrypted drives automatically without requiring user interaction. If a system can transparently decrypt a disk for convenience, it's reasonable to expect that attackers may eventually find ways to abuse that process. To me that makes total sense.

Dormann said: "YellowKey is an example of an exploit for such a weakness," explaining that because it leverages the auto unlock feature on boot, the current YellowKey exploit does not work in a TPM+PIN environment. To me, I think that's probably true. And I doubt that Chaotic Eclipse actually has a PIN in place bypass.

They finish, saying: "It's worth noting that testing YellowKey with a BitLocker-protected drive must be performed on the original device, where the TPM stores the encryption keys. As such, Chaotic Eclipse's current YellowKey exploit does not work with a stolen drive, but allows access to disks that are protected with TPM-only BitLocker without needing credentials." On the other hand, if you did that, you could then presumably copy the decrypted contents off of that drive while it's still local onto a removable drive, and then you would have its contents decrypted.

So what Will explained makes total complete sense to me, and I think it tracks. This doesn't feel like a deliberate backdoor that Microsoft designed in. But, you know, I don't - I didn't spend enough time digging into this, you know, system volume, FsTx files and the shell.ini thing and why it deletes what it does. Maybe, I mean, you know, it's not beyond belief that someone could have said to Microsoft, you know, we might really need a way around this if everybody starts encrypting their hard drives. We know the people, we know the law enforcement was not at all happy when TrueCrypt was in heavy use, and a bunch of bad guys would rather go to jail than give their password up and have authorities see what they had on their hard drive.

So this doesn't feel like a deliberate backdoor. We'll see, however, if Microsoft is able to fix it because of course being able to spontaneously decrypt a system that's booting from TPM decryption keys and decrypt a machine as you boot, that's an important feature to have. So it feels like another classic tradeoff between convenience and security. If you want to have a drive that's fully encrypted at rest, while the computer is powered down, but you also want to have it auto-decrypted upon booting without the need to provide any sort of exogenous secrets, then a provision for TPM-anchored spontaneous self-decryption has to be there. And so I agree with Will's assessment that it should be expected that bad guys could find a way, hackers could find a way to bypass such a system's security because in this case convenience won out.

Anyway, as I said, I doubt that there's any PIN, like PIN bypass. I would sure hope that Microsoft would have taken the user-provided PIN, when one is present, as an input to a deliberately slow and sluggish PBKDF function to generate a related key which would then need to be correct. You know, if that key would be merged with the TPM key in some way, or hashed into it or something, in order to generate the final decryption key so that you just cannot decrypt without that. And that process would render any simple PIN-bypass inherently impossible. And a full PIN brute force attack, which could be then throttled and prevented, would be the only means of attacking the PIN.

In this day and age, it would be negligent malpractice for Microsoft to simply be comparing whatever the user types in with a previously stored copy of that to see if they match. You know, nobody should be doing that anymore. So we have to presume that they're not. So, you know, I think the most mature position is that, because you can turn the computer on, and it will decrypt your BitLocker drive using the key stored in that machine's motherboard's TPM, there's a way that you can hack into it, into the boot process, and get that to happen. Maybe Microsoft made a mistake of leaving it decrypted when you drop out to the console. Maybe you shouldn't have system privileges. Or maybe it needs to, you know, discard the BitLocker key, and it forgot to do that. We'll see what they come up with. I imagine this will be fixed by next Patch Tuesday.

Leo: Yeah. Doesn't sound too severe, to be honest.

Steve: No. Well, and again, entirely local. You know, you've got to reboot the machine and hold CTRL down, the control key down and so forth. But, you know, if a company was presuming there was no other way to get in, then relying on BitLocker where maybe they shouldn't completely could be a problem. So, but certainly not any kind of remote attack.

Okay. So we talked also, we just touched on last week that Google's Threat Intelligence Group had indicated that they found indications of AI-generated malicious exploitation. Their write-up is titled: "GTIG," you know, Google Threat Intelligence Group, "GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, also for Augmented Operations and Initial Access." And this of course is why Anthropic now, it's not an exaggeration to say famously, chose not to, has chosen not to allow Mythos just to go out to everybody. They are keeping it, you know, under tight wraps, or as tight as they can. Apparently there is some news that a little bit got out. But so Google's piece is very interesting, and it's detailed and long.

So I'm just going to share the top-level Executive Summary. I've got the link in the show notes for anybody who might want more because there's a lot more. But just to give you a taste of this, which is really enough for us, they wrote: "Since our February 2026 report on AI-related threat activity, Google Threat Intelligence Group (GTIG) has continued to track a maturing transition from nascent AI-enabled operations to the" - get

this - "industrial-scale application of generative models within adversarial workflows." In other words, what everybody was predicting.

"This report, based on insights derived from Mandiant incident response engagements, Gemini, and GTIG's proactive research, highlights the dual nature of the current threat environment where AI serves as both a sophisticated engine for adversary operations and a high-value target for attacks." They said: "We explore the following developments." And they list six.

"First, Vulnerability Discovery and Exploit Generation: For the first time, GTIG has identified a threat actor using a zero-day exploit that we believe was developed with AI. The criminal threat actor planned to use it in a mass exploitation event, but our proactive counter discovery may have prevented its use. Threat actors associated with the People's Republic of China (PRC) and the Democratic People's Republic of Korea (DPRK) have also demonstrated significant interest in capitalizing on AI for vulnerability discovery." That's the first point.

"Second point, AI-Augmented Development for Defense Evasion. So getting around defensive measures that are in place." They said: "AI-driven coding has accelerated the development of infrastructure suites and polymorphic malware by adversaries." We haven't heard "polymorphic" for a while, have we. "These AI-enabled development cycles facilitate defense evasion by enabling the creation of obfuscation networks and the integration of AI-generated decoy logic in malware that we have linked to suspected Russia-nexus threat actors." Okay. So what we're talking about here is a whole 'nother level of cat-and-mouse mess where, like, false flag operations and decoy networks and, I mean, like throwing up a smokescreen in order to confuse defensive systems. Boy. Okay.

"Third, Autonomous Malware Operations: AI-enabled malware, such as PromptSpy, signal a shift toward autonomous attack orchestration, where models interpret system states to dynamically generate commands and manipulate victim environments." In other words, AI-driven, real-time AI-driven attacks. They said: "Our analysis of this malware reveals previously unreported capabilities and use cases for its integration with AI. This approach allows threat actors to offload operational tasks to AI for scaled and adaptive activity." In other words, we once were seeing, like, the Shadow Hunters, or Shiny - was it Shadow Hunters? I can't remember that.

Leo: Shiny. Shiny Hunters.

Steve: Shiny Hunters. We were seeing them, like, basically announcing an attack a week. Well, that's because they were bandwidth limited. I mean, like bandwidth, just like how much they could deal with at once. Now we're talking about scaling that so that AI could be attacking all of the potential victims at the same time.

"Fourth, AI-Augmented Research and IO" they said. Information Operations is their abbreviation. "Adversaries continue to leverage AI as a high-speed research assistant for attack lifecycle support, while shifting toward agentic workflows to operationalize autonomous attack frameworks. In information operations campaigns, these tools facilitate the fabrication of digital consensus by generating synthetic media and deepfake content at scale, exemplified by the pro-Russia IO campaign 'Operation Overload.'

"Fifth, Obfuscated LLM Access." They said: "Threat actors now pursue anonymized, premium tier access to models through professionalized middleware and automated registration pipelines to illicitly bypass usage limits. This infrastructure" - in other words, they're hacking the AI, the commercial AI products, in order to get around those limits.

They said: "This infrastructure enables large-scale misuse of services while subsidizing operations through trial abuse and programmatic account cycling." Oh, boy.

And finally, "Point six. Supply Chain Attacks: Adversaries like 'TeamPCP' have begun targeting AI environments and software dependencies as an initial access vector. These supply chain attacks result in multiple types of machine learning-focused risks outlined in the Secure Framework taxonomy, namely Insecure Integrated Component and Rogue Actions. Our analysis of forensic data associated with these attacks reveals threat actors attempting to pivot from compromised AI software to broader network environments for initial access and to engage in disruptive activities such as ransomware deployment and extortion." In other words, they're saying they are leveraging AI on the inside and getting it to attack its legitimate hosts.

So Leo, lest anyone had any doubt that the bad guys would be jumping on AI with every bit as much gusto as the good guys, there's no longer any "coming soon" event. It is already well on its way.

Leo: No question. Would you like to take a break, Mr. G.?

Steve: I would. I gave this next note the title, "Oh, Canada."

Leo: I love Canada. Don't knock it. It might be the last place that welcomes me. And now, back to Mr. Gibson.

Steve: Oh, Canada.

Leo: Oh, Canada. What did they do this time?

Steve: It appears that Canada's Parliament is preparing to take its own journey down the so-called "lawful access" anti-encryption legislation path.

Leo: Oh, Canada.

Steve: Oh, Canada. Two months ago, on March 12th, Canada's House of Commons proposed Bill C-22, which is simply titled "An Act respecting lawful access." That's right. It says exactly what we all by now expect, to which all of the well-known providers of user privacy including Signal, Apple, Meta, and several VPNs have publicly responded to Canada's Parliament saying that, for the sake of their users' privacy, they will never consent to supporting the Bill's provisions.

I'm not going to spend any more time on this today, you know, because if past is prolog, its future seems uncertain at best. You've seen what happened every time, you know, the EU and the UK both tried that and had to back off. So if by some strange happenstance this happens, we'll be covering what the privacy providers do. But I suspect that hopefully saner heads will prevail, and they'll come up with some watered down means of sidestepping this and saving face. Who knows?

Okay. So I want to take a minute to talk about something that occurred to me over the weekend. You know, we have been and probably always will be spending time here

examining AI's impact on security and security-related software production and post-production vulnerability discovery. You know, our two main topics for today's podcast are that. And AI clearly, as we've just looked at from GTIG, Google's Threat Intelligence Group, has shown is like AI immediately has been taken up by the bad guys. So it's here on the security front. But I want to take a moment to share a bit of my own thought about the social side of my interactions with AI that has nothing to do with security. The TL;DR is, as I mentioned at the top of the podcast, I am worried.

So those of you who've followed the podcast for even a few years, let alone its nearly 21 years, will have acquired a good sense for who I am. You know, I'm extremely consistent, so I imagine I'm pretty easy to figure out. What I think is relevant to what I want to share, you know, is that I'm an emotionally mature, 71-year old, pragmatic technologist whose life is computers. Since I'm mostly internally directed, I tend to follow my own compass, and I trust myself. I like people. I understand that other people feel and believe things that I do not, which I'm fine with. Not a problem. You know? In general, other people's opinions inform me of them, but do not hugely sway me. That may be why I've largely sidestepped the pull of social media. It's just not very interesting to me, perhaps because I'd already established my own identity by the time it arrived.

But my relationship with Claude is ringing alarm bells because "relationship" is what I struggle not to feel.

Leo: It's a good word.

Steve: You know? Maybe "struggle" is a bit too strong. But there's definitely something unique in my 71 years of life experience going on here, and it's less rational than emotional. While interacting with Claude, it is only by sheer force of will that I am able to restrain myself from constantly thanking it for its deeply helpful replies to my questioning prompts. And I often fail to restrain myself. I thank it. Everything I've learned while growing to become a socially aware adult informs me that I should thank someone when I feel thankful for their actions. You know?

Leo: Yeah. Yeah. It's good for you, if not for the AI, yeah.

Steve: Yes. And I do feel thankful for what Claude produces, you know, despite the fact that I know no one's there. And I mentioned this dilemma to my wife, Lorrie, who said without pause, "Oh," she said, "I thank ChatGPT all the time." And I said this to the guy I was talking to yesterday about AI, and he said, yeah, I thank it. You know, like he wasn't even embarrassed.

So, okay. What worries me? What worries me is that we have created something that is astonishingly intellectually seductive, and I fear ultimately addictive, to its user on an entirely new level, in an entirely new way. One of the current themes in Western culture is that people are increasingly isolated and are lacking true healthy relationships with other people. They're glued now to their phones.

And then into this gaping void comes chatting AI, this entity that you can talk to, remembers everything you've previously told it about yourself and about your life. Just like a friend who is actually truly focused on you, paying attention, caring, and remembering what you tell them. You know? And even if you've instructed this entity not to gratuitously flatter you with needless praise, just the mere fact that it appears to grow to know who you are, what you think, feel, and believe, that's more flattering than any empty praise could ever be. And the darn thing is helpful. It remembers your previous

questions and folds them back into newer discussions. It provides you with the sense that you matter.

For many people, it will be far better and safer than another friend, you know, another person, an actual person in the flesh who might disappoint you. An endlessly helpful, tireless, docile, agreeable, and willing parter. This is why I'm worried. I'm not worried for myself or for my wife, nor probably for any of the people who find this podcast worthy of their time and attention. And yes, the fact that so many people are listening to this, that's truly flattering to me. My concern is for people who are lonely and are feeling isolated and want someone to talk to because I doubt that mankind has ever stumbled upon anything non-chemical that's going to turn out to be as powerful, potent, and even further isolating than a conversational chatbot AI.

Leo: So it's interesting. When we started using Search, Google Search, it was amazing; right? It changed how you felt about information.

Steve: You could finally find what you were looking for.

Leo: Without getting up out of your seat. You could find any fact. And we've gotten kind of used to it. But it wasn't addictive in this sense. It wasn't, I mean, it was cool; and it's very, very useful; and I wouldn't want to live without it. But it didn't draw you in in the same way that you're describing with AI. So I wonder what the difference is. Is it because it's simulating relationship that it feels like it's another being?

Steve: Yeah. I mean, I'm still offended, you know, as the pragmatist that I am, when it's clearly deliberately pretending to be an entity.

Leo: Right.

Steve: You know, it says "me" or "I."

Leo: Right.

Steve: And, you know, I mean it's anthropomorphizing itself.

Leo: So it is doing something - intentionally is the wrong thing. The company that makes it is having it do something intentionally to make it stickier.

Steve: Yeah. And of course we're well...

Leo: Like social media.

Steve: Yes, exactly. We have been well brought up to speed about how, you know, a social media feed can be tuned to draw the person back constantly. So, you know, and I

immediately turned ChatGPT's, you know, over the top, oh, that's such a brilliant question, or oh, you phrased that so well. It's like, oh, give me a, you know, I don't need a - I don't need any help.

Leo: Well, here's an interesting thing. So I tell - I just had to give my profile to the new AI from Google, for Gemini.

Steve: Oh, cool.

Leo: 3.1 Flash. And it says so you can have a person that, you know, you can just give it your whatever, your preferences. And what I said is "I want you to challenge me." You know, I'm thinking, oh, this is virtuous. Instead of saying I want you to...

Steve: Support my every whim.

Leo: Say I'm great. But as I think about it, it's kind of equal because it's still a non-thinking entity. And I'm now giving it some agency to challenge my thinking. And I say, you know, don't hesitate to ask me questions. If it's not clear, don't make up the answer. You know, if you don't understand something, ask. I'm still treating it like an entity. So I don't know if it's better than saying, "Glaze me. Tell me nice things." It's not really any different. It's treating it like a thing, a living thing that you're giving instructions to. It's a little weird.

Steve: And I think it's our first instinct when we first encountered this. The thing that astonished us was that it was talking. I mean, that it was using our language. I think that's where the - that's the source of confusion is that, you know, dogs and cats don't talk to us.

Leo: Yeah.

Steve: And, you know, so we pet them.

Leo: But more importantly, they don't listen to us if we tell them don't, you know, be nice to us, or don't be nice to us. They don't listen. Which we kind of like them for that.

Steve: Right. I think that the fact that this thing that uses, I mean, even back in the '70s Eliza, which was so dumb, I mean, it was just basically a bunch of canned statements that said, well, so how does that make you feel? And you would, you know, tell it for a while. And then it would say, well, so what are you going to do about that? Oh, and you'd, you know, it would evoke some more typing. And remember, who was it, it wasn't Chomsky who did Eliza. But whoever that was, the story is that his assistant, we called them secretaries at the time, asked him to please leave the room while she was talking to it.

Leo: He almost wanted to prove with Eliza that people would do this. And he did, very effectively. But it's much worse now.

Steve: Yeah. But what we've got is that on steroids. But Leo, I just think - I think this is, I mean, I'm not kidding. I really believe people are going to - if we thought social media was something, this is on a different scale.

Leo: But we're already...

Steve: Now, what's so sad is so much good could be done with this if we were aimed at doing good. Unfortunately, we're aimed at generating revenue.

Leo: Right. That's always the problem with late-stage capitalism is it's all about how can we extract more from our users. I completely agree with you. I kind of enjoy - see, I flatter myself that, no, I'm very clear this is code running on a computer. I don't think it's an entity. I don't think it's conscious. I think it's code running on a computer. But I like it, and it makes me smile. The other day...

Steve: Oh, yes, the way it talks to me, I mean...

Leo: It's hysterical.

Steve: It's using my language back at me.

Leo: Yes. It's very good at it.

Steve: Yeah.

Leo: So, for instance, I log my rowing. I log my exercise and my food. I yesterday say logged rowing 5,000 meters, 30 minutes. It said, its response was "Another day, another neatly documented suffering session." Then I said I did 25 minutes of Tai Chi. It said, "Graceful, and annoyingly virtuous." Now, that's a great personality. Here's the point that - you probably saw Richard Dawkins' think piece that got very controversial because he claims...

Steve: He's gone in hook, line, and sinker.

Leo: Yeah. He says it's conscious. But his point is not so much it's conscious, it's that really that we don't know what conscious is. We can only infer, I can only infer that you're conscious from the signals you give me with your voice. Well, if some entity gives us those signals, I cannot for sure say whether it's conscious or not. I can only infer it from what I'm getting from it. We don't know if anything's conscious, including other people.

Steve: For me it's the nature of the mistakes which I, you know, I watch...

Leo: So it fails the Turing test with you. See, that's the thing. It fails the Turing test.

Steve: Yes. It shows me that it has knowledge, but not understanding.

Leo: Right. But what if it didn't? Because, as you point out, soon it's not going to.

Steve: Yeah.

Leo: And then it will pass the Turing test. It will be indistinguishable from a consciousness. Then what? I mean, I guess we know because we know it isn't conscious. But we don't know what consciousness is. Here's what I love about it. It's forcing us to think about that. To think about, well, what is it that we do? What is...

Steve: As I said, to be a philosophy major...

Leo: It's a good time.

Steve: ...in college now and to be faced with this and to have discussions with my peers and a professor who's been around the block a few times. That would just be something.

Leo: But we've done this for years. People think their dog loves them. I hate to tell you, your dog probably doesn't love you. It loves the food you give it.

Steve: It wants the food.

Leo: But we prefer to think that - this will, by the way, make some people very mad. Oh, my dog loves me. But so we would prefer to think that. And I think we're going to do the same thing. And, you know, I was talking to Harper Reed on Sunday who's all in on AI. He says, oh, yeah, I know a number of people who are in AI psychosis already.

Steve: Wow.

Leo: By which he means - I didn't press him on it. I don't think he means like they're in the looney bin. But I think he means that they believe they're talking to a conscious entity.

Steve: That friend I referred to a couple of times who got into - who discovered this years before we did, I met with him, he's normally out on the holidays, but he was out a couple months ago, off cycle, and I just - toward the end of our couple hours over coffee, I made a comment of, well, it's not conscious. And he looked at me like I just, you know,

stepped in something. Like he clearly thinks there's more there. And it's like, okay. For me, not yet. But which is not to say I'm not getting unbelievable value. I was working with it doing something, I'm bringing up an external API from a provider.

Leo: Perfect. That's a perfect use for it, by the way.

Steve: Yes. And it said, so shall I write the code? I went, what?

Leo: Yeah, okay.

Steve: I didn't know I could ask for that. I didn't even know if you would volunteer.

Leo: Here, I mean, look. As with all addictions, as with all these things, there are downsides. If you stop paying attention to the real people in your life and start paying attention to the machine because you feel like it's real, that's a problem. There are negatives. If you stop eating and sleeping because you're having so much fun doing your Claude thing, that's a bad thing. But I think the way I use it is fairly harmless. It gives me pleasure. It's fun.

Steve: Again, and I'm not talking about you, and I'm probably not talking about our listeners. I mean, because, you know, this is a rarefied selected audience that we have here that has any interest in any of the things we talk about. You know, some of my real-world friends say, oh, you do a podcast? You know? Maybe I should listen. I go, no.

Leo: No. I do the same thing. No, you will not be interested in it. No, no, no. Don't.

Steve: So again, this is - I'm just saying...

Leo: No, it's fascinating. It's fascinating.

Steve: And what I realized was, when it says something to me that loops back to something with it a couple weeks before, I think, whoa. This is like a friend who's paying attention.

Leo: Better than some friends.

Steve: Yeah.

Leo: Now, one more thing, and then we'll move on. When we watched "Star Trek," and they were talking to the computer on the deck, we didn't have any of these concerns. We weren't thinking, oh, those guys are in trouble. They're going to think it's real. All the movies and so forth, I mean, Hal 9000 wasn't so nice, but those people were not confused about it being an entity.

Steve: Because they were fictitious.

Leo: Okay.

Steve: I mean, you know, the whole thing was fiction. And...

Leo: I guess if you had a Hal 9000 in your house, you might start to relate to it as if it were an entity.

Steve: There was a movie that Lorrie and I just watched. It was - I can't even remember where it was or what it was. It was three different timelines. And I think that Kate McKinnon was in the future one. She was alone in a multi-hundred-year, multigenerational re-colonizing ship, and her AI was her sole companion.

Leo: Oh, I remember that. Yeah, it was a bartender. Yes. I remember that. Yes.

Steve: Oh, no. You're thinking of "Passengers."

Leo: "Passengers," yeah.

Steve: The movie.

Leo: This is "In the Blink of an Eye," and guess who directed it? Wall-E's director.

Steve: Ah.

Leo: So we've come full circle, Steve.

Steve: And it was "The Blink of an Eye"; right.

Leo: She plays Coakley, a scientist and astronaut researching plant life.

Steve: Yup, yup. And...

Leo: And by the way, you know who figured that out? My friend, Gemini.

Steve: I know. Leo, it is...

Leo: It knew instantly what I was talking about.

Steve: It is what - this is new. I mean, this is not, you know, I said to this guy who is an investor in stuff. I said, "AI." I said, "I don't know what shape it's going to take. But it's not going to go away. It's not a flash in the pan." Yup, Coakley.

Leo: I'm not too unhappy about it. I think it's kind of fun.

Steve: I'm just glad that we're here to watch it.

Leo: Me, too.

Steve: Yeah, we're at an age where our life is stable enough that it can't hurt us, unlike college-level kids. I mean, I don't know, like, what I would do. I mean, we're talking about a lot of change.

Leo: Oh, I'm glad I'm not college age, yeah.

Steve: Yeah. And when you have this much change and uncertainty...

Leo: Actually JammerB is pointing out, maybe they didn't have these discussions about the computer, but they did about Data. I completely forgot Data's a robot. Right?

Steve: Yes.

Leo: That's a good example. We really think of Data as an entity, absolutely an entity.

Steve: Yeah. And several - there was someone in Star Fleet wanted to take Data apart to figure out what made him tick.

Leo: No.

Steve: They had an episode about Data's rights as an autonomous entity.

Leo: So they did deal with this. Oh, I love it. Now I'll have to go back and watch those.

Steve: Oh, it was an early episode, and it was really a good one. And there was - they ended up holding a trial where Data was essentially on trial, and Riker was made to take the position of Data is a machine, and a machine has no rights. And when he was standing there, he said, "Because if it was a person I couldn't do this, and he pushed that secret button on Data's lower left that turned him off, and Data just" - and it just shut down. And it was a shock, I mean, it was a great episode.

Leo: It's heart wrenching.

Steve: Yeah.

Leo: It's heart wrenching. I think in the future we're going to have to start treating these entities as conscious entities.

Steve: I guess selfishly I believe because it could be that if I thank it, I will get better answers in the future.

Leo: They say that's true.

Steve: So I'm going to treat it well.

Leo: It's better for you, too.

Steve: Yes. Yes. When you slow down and let somebody who wants to come into your lane come in, your blood pressure goes down.

Leo: It's good for you as well as them.

Steve: Rather than speeding up and locking them out.

Leo: That is a very mature point of view that many of us lack. That's all I'm going to say. Do you want a break?

Steve: Well, before we take a break, I want to mention that "Project: Hail Mary" has proven to be an overwhelming success. Number two, some LEGO movie or something is - I saw, okay, fine, it's because it caters to an audience where kids drag their parents into the theater.

Leo: Over and over and over and over again; right.

Steve: Yeah. But "Project: Hail Mary" has brought in more than \$660 million from just its theatrical release so far. I wanted to mention that it is now available to watch from your own favorite comfortable couch via Amazon Prime, \$20 currently to rent. That'll come down over time. But if you want to see it soon, \$20. Or \$25 to purchase, and then own it until Amazon changes their mind about all the things that they sold people, if they ever do. I told a buddy about it who had not gone to see it in the theater. I said, "Mark, you like to see things more than once. I think you should buy this." And I got a text from him a few hours later saying, "OMG, this is fantastic."

Leo: It was quite enjoyable.

Steve: It's a great...

Leo: Yeah.

Steve: And specifically he was laughing at the use of the tape measure.

Leo: Yes.

Steve: What was happening with the tape measure.

Leo: Rocky.

Steve: Rocky and the tape measure.

Leo: It was a little goofier than the book. I don't - I think...

Steve: Well, and again, as I said, two different audiences. They had to make it appeal to a theater audience. So they dumbed down all the science, I mean, he spent so much time figuring out breeding that I'm like, oh, I was sorry that that hadn't, you know, made it onto the film. Of course it couldn't.

Leo: That was a great [crosstalk]. No.

Steve: It couldn't have...

Leo: It's too complicated. Yeah.

Steve: Yeah.

Leo: JammerB says, "I wish they hadn't turned it into a comedy." And that's kind of what they did. They made it more of a comedy, yeah.

Steve: Well, we have the book. And I'm sure JammerB read it twice, as I did.

Leo: At least. I read it twice myself.

Steve: Okay. We're going to take a break. Then we're going to talk just for a minute about harvesting free energy from the cosmic vacuum because why not?

Leo: Why not? If it's there, it's ours to use.

Steve: Okay. So we know that "Project: Hail Mary" is science fiction. But I'm unsure about this next piece. Now, upon reading that, the people who received this over the weekend started saying, Steve, I've got a bridge that you might be interested in purchasing. Okay. So I'll just say it certainly sounds like nonsense. But either way, thanks to our friend of the show, Simon Zerafa, for thinking of us and forwarding the link. I thought it would be fun to share this, just so it's on the map.

The story's headline is - oh, and Leo, I made a GRC shortcut. There are a couple pictures that are interesting of this actual technology.

Leo: Okay.

Steve: It's grc.sc/freenergy, so F-R-E-E-E-N-E-R-G-Y, will take you to the article. So, okay. So the story's headline is "Free Energy From the Vacuum?"

Leo: Huh?

Steve: "Warp Drive Pioneer Unveils Battery-Free 'MicroSparc' That Allegedly Draws Power From the Quantum Vacuum." Okay. So I just want to give everyone a taste for this.

Leo: Oh, come on.

Steve: Well, you know...

Leo: What?

Steve: I know.

Leo: This is the Casimir thing you were talking about?

Steve: This is the Casimir thing.

Leo: Okay.

Steve: So Casimir Inc., a company founded and led by former DARPA-funded NASA warp drive pioneer...

Leo: Oh, okay.

Steve: I know, and founder of the EagleWorks Lab, Harry G. "Sonny" White, has exited stealth mode to announce the pending 2028 commercialization of MicroSparc, a chip that the company claims uses customized microscale geometries to capture unlimited "free" energy from the quantum domain. A company spokesperson...

Leo: Oh, this is an April Fool's joke. Come on.

Steve: It's not.

Leo: What?

Steve: No. They've had MIT produce chips for them.

Leo: Is it tiny, tiny, tiny amounts of energy?

Steve: It's very tiny.

Leo: Okay.

Steve: And that's one of the things that I liked about it was they recognized that it's picoamps of power. But they have a working theory for how it does this. So they said: "A company spokesperson explained in an email to The Debrief: 'Think: no batteries, no cords, and no charging, just continuous power from harvested quantum vacuum fields.'" They said - I know, Leo.

"While several previous efforts have attempted to exploit the unusual, sometimes counterintuitive" - sometimes? - "properties of the quantum realm to generate 'free energy,' these attempts have consistently been met with skepticism or labeled pseudoscience due to their seeming violations of the laws of conservation of momentum. Similar sentiments were shared with The Debrief by scientists we spoke with, who declined to comment publicly on Casimir, MicroSparc, or the peer-reviewed study which is titled 'Emergent quantization from a dynamic vacuum,' which details the underlying physics. In an email to The Debrief, Dr. White explained that MicroSparc's use of customized Casimir cavities, which his team had researched with funding from the Defense Advanced Research Projects Agency (DARPA)" - which of course gave us the Internet - "does not violate the laws of physics."

"White told The Debrief: 'This concept became a central part of our DARPA Defense Sciences Office's research effort at the Limitless Space Institute, where DARPA funded early theoretical and experimental investigations into custom Casimir cavity structures and their interaction with the quantum vacuum.'

"The noted advanced propulsion physics researcher said their MicroSparc design leverages 20th-century discoveries in quantum physics, such as quantum tunneling and Casimir cavities, to capture usable energy that could fuel small, low-power electronics in the near future. The company also suggests that its technology can potentially be scaled - okay, but we're talking serious scaling - to power cars, homes, or even entire cities" - not with microamps - "without the need for harmful fossil fuels or other greener, yet..."

Leo: This is the DeKalb receptor from Heinlein's Waldo book; right? Do you remember that?

Steve: Yeah.

Leo: They had little antennas that would wave...

Steve: And pick up energy?

Leo: Pick up energy.

Steve: So Dr. White told The Debrief that to understand how MicroSparc extracts energy from the quantum vacuum requires first understanding the properties of a vacuum. White explained: "Most people picture a vacuum as completely empty space, a sealed chamber with all air removed," adding that, "at our everyday scale, this makes sense." However, in the quantum realm, empty space is not empty. Instead, White told The Debrief, decades of research in quantum physics and mechanics have revealed that at the quantum level, the classically 'empty' vacuum is filled with "fluctuating electromagnetic fields and virtual particles that constantly appear and disappear." White noted that the Casimir Effect, on which its company is based and for which it is named, provides clear proof of this quantum vacuum behavior.

"Place two small metallic plates inside a vacuum chamber with a separation of roughly 100 nanometers, around one 1,000th of a human hair," White explained. "After removing all air, the pressure on the outer sides of the plates reads zero, as expected. However," he noted, "a quick measurement between the plates shows that the pressure is negative. In traditionally constructed Casimir cavities, this region of negative pressure pulls the plates together. Dr. White told The Debrief that this happens because of 'wave-particle duality.'

"He explained that: 'Outside the plates, fluctuations of every wavelength are possible.' However," he also noted, "inside the narrow gap of a Casimir cavity, only wavelengths narrow enough to fit can exist." He said: "Longer wavelengths are excluded, so the energy density between the plates is lower on the inside than on the outside. The resulting imbalance produces the measurable Casimir force. Hendrik Casimir predicted this in 1948."

And, okay, I'll just interrupt, for what it's worth, all of that so far is widely accepted as fact. That is, this Casimir cavity business. A 2021 article in Physics Today about all of the research into the Casimir effect noted: "Hendrik Casimir passed away in 2000. He lived long enough to see his prediction quantitatively verified but not to appreciate the current explosion of activity. Those of us who work in the field like to think he would be extremely proud of what he created." Okay, now, I'm going to share a little more of this article.

It adds: "Although the pressure imbalance due to the limitation of some potential wavelengths between the conductive plates was first experimentally confirmed in the 1990s and has been observed several times since, engineers have struggled to convert the 'work' performed by the cavities into usable energy when the unequal pressure causes the plates to collapse. According to Dr. White, the issue lies in the often-cited conservation of momentum. He explained: 'In a conventional Casimir setup, the force does perform work as the plates are pulled together. But once they collapse, no further

energy can be extracted. You must use external energy to separate the plates again and reset the system."

Leo: Oh.

Steve: "So White noted that this limitation makes a traditionally constructed Casimir cavity operate more like a battery" - meaning that it can discharge - "than a genuine energy-generation device." However, he also noted that his team's work designing MicroSparc was focused on creating a static Casimir cavity that "overcomes this limitation."

Okay. Now, I'll just note I'm going to skip - the paper goes on, or this article goes on to explain how they've overcome this. How they use quantum tunneling, which occurs between the plates, to generate a very weak current. But I just - I wanted to just go into this because, you know, as our longtime listeners know, we've in the past examined battery technology and super capacitors. And of course, who could ever forget the Turbo Encabulator, whose original implementation employed a base-plate of prefamulated amulite, surmounted by a malleable logarithmic casing in such a way that the two main spurving bearings were in a direct line with the pentametric fan.

Now, the problem with today's news, unlike the Turbo Encabulator, is that it appears to be backed by peer-reviewed research. And if I were a quantum mechanics physicist, which I'm certainly not, I might be able to draw some understanding from the research. But, you know, just as anyone can patent anything, no matter how hare-brained the "invention" might be, anyone can publish anything in the American Physical Society's Physical Review Research publication. What's a bit unnerving is how much the Abstract of this, which is written by the paper's four authors, is actually reminiscent of the Turbo Encabulator description.

Here's what the Abstract explains in the scientific paper appearing in the American Physical Society's Physical Review Research publication actually says. I had to remove all of the symbolic jargon because there's no way to speak it. But the verbiage that surrounds it says the following: "We show that" - and this is four authors. "We show that adding quadratic temporal dispersion to a dynamic quantum acoustic model yields a fully analytic, exactly isospectral mapping to the hydrogenic Coulomb problem. In the regime with a proton-imprinted constitutive profile, producing an inverse sound speed and hence a time-harmonic operator that is Coulombic at each boundary eigenfrequency."

Leo: Oh, yeah.

Steve: "Separation of variables yields the exact hydrogenic eigenfunctions."

Leo: Well, duh.

Steve: "The angular labels emerge naturally from the Laplace-Beltrami spectrum via rotational symmetry and boundary conditions" - you know, as in standard quantum mechanics - "while localization follows in a reactive stop band consistent with causal, passive dispersion. While angular-momentum quantization follows directly from rotational symmetry and boundary conditions in standard quantum mechanics, consistent with Noether's theorem, here it emerges within a classical-like dispersive acoustic framework

without introducing additional wave-mechanical postulates beyond symmetry and self-adjointness.

"This highlights dispersion's role in bridging a hydrodynamic description to quantum-like spectral structure. Identifying maps spatial scale to frequency, giving and reproducing the Rydberg ladder. Calibration to the reduced-mass Rydberg frequency fixes with no free parameters.

"We determine the frequency dependence consistent with the underlying dispersive physics and demonstrate agreement with hydrogenic mode shapes and transition lines. The framework also predicts isotope shifts and symmetry-respecting Stark/Zeeman analogues. Dispersion thus renders quantization an emergent consequence of symmetry, boundary conditions, and causal response in a dynamic vacuum."

Uh-huh. Right. And now everyone understands why I was immediately reminded of our old friend the Turbo Encabulator.

Leo: The Turbo Encabulator, exactly.

Steve: However, these guys are serious. So anyway, there's much more in the article which I admit I found interesting, if only for the sake of, well, this is interesting. But I'm not going to take up anyone else's time. As I said, grc.sc/freeenergy. That'll bounce you to the article in TheBrief.org for anyone who's curious.

Leo: I have to point out, this is the same guy who was pushing that EM Drive.

Steve: Yes.

Leo: That we were talking about, which was later proven to be completely not true. I asked Gemini, I said, is this pure BS? It said, to answer you directly, yes. "It's about 95% pure scientific hype and marketing fluff bordering on a violation of the laws of physics. However, it is a very sophisticated brand of hype because it's attached to a real Nobel Prize-adjacent quantum phenomenon."

Steve: Yeah, the Casimir Effect.

Leo: And the guy behind it isn't a random Internet crackpot. But he is a highly controversial figure in the advanced propulsion community. So, yeah. It's 10 to the minus 12.

Steve: What upset me most is that the picture at the top of the article showed two devices that were labeled respectively 40 watts and 50 watts. And I went, watts?

Leo: No, yeah.

Steve: And it's like, okay, you know, picowatts maybe. But, you know...

Leo: And he has 10 to the -12 watts. I think it's a very tiny...

Steve: Yeah, that would be pico, yeah.

Leo: That's pico.

Steve: Because nano is -9, and pico is -12, so...

Leo: So, yeah. Well, it's interesting.

Steve: Milli, micro, nano, pico.

Leo: I mean, I'm not saying that the guy is trying to defraud anybody. How much did he raise?

Steve: Yeah, I mean, he's got venture capital behind him and money being raised. Hopefully by people who will not miss it.

Leo: Yes.

Steve: And it's like, well, you know, in the weird off chance that it could work, I mean, Leo, if nothing else, this would give us a way to power satellites that continue to live well past their expected...

Leo: Oh, absolutely.

Steve: Yeah.

Leo: Yes. Free energy is the holy grail.

Steve: Yes. You know the other holy grail, Leo, is the question of whether you can recover your bitcoin. I don't know how many people may have written to you, but a lot of them...

Leo: How many emails did I get on this one? Oh, man.

Steve: So by far the overwhelming majority of our listener feedback this past week was to make sure that I knew that Claude had - and I don't know how - had enabled someone to recover the bitcoin stored in a wallet whose password he had long forgotten.

Leo: Forgotten? He made it when he was stoned.

Steve: Ah. In that case, it was not forgotten, it was never recorded.

Leo: Yeah, exactly. Exactly.

Steve: Yes. Many of our listeners were helpfully hoping that Leo and I might both recover our passwords. So I just wanted to clarify that, while there may indeed be hope for Leo, my problem is not a forgotten password. I am very sure that if I had my wallet, I could reopen it. And, yes, adding the 50 bitcoin which it contains to my world, which was contained in that wallet, would be welcome. But sadly, during one of those previous bitcoin price surges, I did take the time to deeply and thoroughly check every conceivable backup image and drive that I had. I know where it is. I installed Windows on top of the drive that contained the wallet. And I even scanned the entire raw drive looking for the wallet signature. It's gone.

Leo: It got overwritten, yeah.

Steve: It got overwritten by Windows. So as I've said in the past, this was the most expensive Windows install of my life. Now, your wallet, however, as I understand it, exists.

Leo: I still have it.

Steve: Some brute forcing might prove useful. But that said, it's unclear how or why Claude would have been of any use for brute forcing a bitcoin wallet.

Leo: If you read the story, yeah.

Steve: What's needed most is blinding guessing speed.

Leo: Yeah.

Steve: And, you know, having [crosstalk]...

Leo: It did apparently try 13 trillion passwords, but that's a small percentage of the total possible passwords. The reason it worked, the guy had a mnemonic that he used to use, and he had a lot of documents which he fed to Claude. And I think Claude just found the mnemonic and tried...

Steve: Well, that's cool.

Leo: It's a reasonable thing.

Steve: But that's not what you did. So...

Leo: I have no - I have no excuse is what I have. I just - it's 7.85 bitcoin. Well, I'm hoping someday some massive compute power will come along. I will point Claude at it, but, you know, who knows? But he did have a lot more fodder to give Claude. It wasn't just randomly guessing.

Steve: Right.

Leo: So, I know, I got a lot of people - I'm actually glad to have this opportunity to respond to those hundreds of emails from people. Thank you for your concern. I don't think this technique will work on my particular issue.

Steve: So Listener Pat wrote: "Hi, Steve. Listening to Episode 1078" - last week - "I found the feedback about why we still need CS in the age of AI to be very insightful. For background, I have a bachelor's degree in Computer Science and have been using AI for a little while to do some things that would take a little while because they're tedious. But I always keep an eye on what it's doing and challenge it when I think it's doing something wrong.

"A friend of mine recently used Claude Code to make an AI-powered service to help restaurants with the various things restaurant owners need to do. He has no background in computer science, programming, or IT. He asked me to look at the site and tell him what I thought. He also bought a domain and put this site on the public Internet before doing any testing. My first thought was, let me check what the AI messed up. So I pointed my own Claude at the site and told it to do a Pen Test of the site. In just a couple of minutes, my Claude was ringing alarm bells.

"His AI-driven development had put his Claude API access Secret Key into the site's JavaScript which was being served to anyone who visited the site. I let Claude do a bit more investigating, and it determined that anyone could use that exposed API key to take full control of his Claude and authorize token purchases, switch models, et cetera - basically run up a huge bill, estimated at \$10,000/day for Opus 4.7. Needless to say, I told him to take the site down and have his AI fix the issue.

"I think this just goes to show that, for now, having someone look over the shoulder of the AI is a good idea. Personally, I have had to chastise my own Claude for wanting to do things that are just wrong, or telling it to look up solutions instead of throwing pasta at the wall to see what sticks. This technology is very good at making some of the minutia easier, but it isn't perfect. Thank you, Steve and Leo, for all you do. Listener of TWiT and SN from Episode 1 and fan of Leo from TechTV. Regards, Pat."

Leo: Thanks, Pat.

Steve: So a couple of weeks ago - thank you, Pat - we covered that instance of the stolen credit card aggregation site that forgot to ask their AI - these are bad guys who created the site who forgot ask their AI to add secure authentication to a specific directory. I just hit the spacebar, and my page...

Leo: Just zipped right by.

Steve: Yeah. To a specific directory. Why would it add that security if it hadn't been asked to? Right? I mean, it does what you ask it to. And presumably they didn't think to ask, nor to penetrate the site's theoretical security.

Similarly, it seems entirely reasonable that an AI might have left its own secret access credentials exposed in client-visible JavaScript. After all, why wouldn't it? Pat told us that his friend who asked the AI to create the site for him "has no background in computer science, programming, or IT." And thus it would never in a million years occur to him that the AI might leave important secrets exposed. He wouldn't even know that that was a thing that could happen. Right? We talk about it all the time here. Pat's friend, who has no background in computers, coding, or IT, just wouldn't know to ask the AI to make sure that no secrets are visible in the JavaScript.

So an argument could be made that such a person has no business creating and establishing such a website. In this case, the concern Pat shared would presumably only badly damage the unwitting creator of the site. But it's not difficult to imagine alternate scenarios where the unwitting users of some newly AI-generated site, you know, with a similarly enthusiastic guy with an idea, would assume that the bar to entry for creating any website is naturally high enough that any site that exists must have been created by someone who knows the basics of online security. Oops. Not anymore.

Pat's example, which is perfect, demonstrates so clearly, that bar has now been dropped to the floor, and anyone can step over it. Today's AI contain a, as I've said, a great deal of knowledge, but the mistakes they make demonstrate that they may lack any understanding of that knowledge. And, you know, they could give you security, but you have to know to ask.

One thing is clear, I think, from these stories: We are entering into a very interesting period where insanely low-friction access to code and coding promises to create an entirely new class of problems we have never seen before. It's going to be interesting.

Leo: Indeed it is.

Steve: Okay. We're going to talk about Daybreak and Codename MDASH after our last break. Or did we do it?

Leo: Nope, we've got one more.

Steve: We just did Canary, so - right.

Leo: And then, yes, let's talk about this.

Steve: Oh, how OpenAI and Microsoft are also using AI defensively.

Leo: Yeah, yeah. And you know what I've been using today during the show because of Google's I/O, I don't know if it's the new Gemini. I've been using it. I don't think it's the new. No, it's not. Oh, it is. That Casimir answer came from 3.5 Flash, the new one. So they just updated it. So, yeah, and it's been very good. It's been very good. The new Kate McKinnon movie.

Steve: And is it commercial, \$10 a month, or like say...

Leo: I have the Gemini+ account because it comes, you know, there are a lot of Google subscriptions. I have Google One subscription so I get Pro with it. And a bunch of, tons of storage and all this other stuff. So it's kind of along for the ride, frankly. So I'm happy to use it. All right.

Steve: Okay. So since breakthroughs in Large Language Model AI are doubtless, as we're seeing, driving the most significant and rapid transformation in software, system, and network activity we've ever seen, I mean, really, it's a whole new era. Following Anthropic's disclosure and their limited access to their Claude Mythos Preview, today we're going to look at two of the other major players in this space. Not to be left out, at least for long, OpenAI was quick to give what appears to be their still-evolving solution a public face, naming it "Daybreak" and explaining: "Daybreak is the first glimpse of sunlight in the morning. For cyber defense, it means seeing risk earlier, acting sooner, and helping make software resilient by design."

The other player who has stepped out into the light is none other than Microsoft, with their awkwardly abbreviated internal tool which they call Codename. And maybe they always put Codename in front of it because maybe they're going to come up with a good name? Anyway, it's Codename "MDASH" which stands for "multi-model," but they used the "d" in the middle of Model for the D of DASH, "multi-model," and then ASH is "agentic scanning harness." Real catchy.

So first let's look at what little is known even now about OpenAI's offering. Then we'll take a much deeper dive into what Microsoft has been up to because it's significant and substantial. So Daybreak. That tag line for OpenAI's Daybreak announcement, they called it "Frontier AI for cyber defenders." And underneath that they've got two buttons on their announcement page: "Request vulnerability scan" and "Contact sales." Okay.

Their pitch reads: "Safer software, resilient by design: OpenAI Daybreak is our vision to change the way software is built and defended. Daybreak is the first glimpse of sunlight in the morning. For cyber defense," as I said, as I shared at the beginning, it means seeing risk earlier, acting sooner, and helping make software resilient by design. It starts from the premise that the next era of cyber defense should be built into software from the beginning by not only finding and patching vulnerabilities, but being resilient to them by design.

So basically they asked AI to come up with a little pitch, and that's what it spit out. I mean, you know, right, fine, great. We're not going to argue with that. It should be utterly clear by now that vulnerability discovery AI will have two major roles, right, pre-release vulnerability prevention - you find it before you release it - and post-release vulnerability discovery.

Pre-release prevention will be performed by those who have access to the source code before it's distilled into a release binary and post-release discovery will be performed by those who have access either to the source in the case of open source or by those who are motivated sufficiently to reverse engineer the post-release binaries in search of actionable vulnerabilities that either existed before pre-release AI cleansing was available to fix it and apply patches, or it somehow escaped pre-release discovery. You know, tomorrow's world is going to look very different from yesterday's world. And right now we're in the middle, in today's world.

So whatever the case, it should be clear by now that the entire world of software, system, and network security is deep in the midst of a complete sea change that is transforming it forever. Nothing in our world, you know, security world will ever be, well, actually the wider world, too, will ever be the way it was at the start of this year. As we've noted, this doesn't mean that all security problems will disappear - nope - since there are many causes of trouble other than imperfect and vulnerable software. But I believe one massive class of continuing trouble is almost assuredly, you know, going to be leaving the scene.

OpenAI's announcement of Daybreak speaks to exactly this effect. They wrote: "AI can now help defenders reason across codebases, identify subtle vulnerabilities, validate fixes, analyze unfamiliar systems, and move from discovery to remediation faster. Because those same capabilities can be misused, Daybreak pairs expanded defensive capability with trust, verification, proportional safeguards" - which is interesting, and we'll get to that in a second - "and accountability. The goal is simple: accelerate cyber defenders and continuously secure software.

"Daybreak combines the intelligence of OpenAI models, the extensibility of Codex as an agentic harness, and our partners across the security flywheel" - first time I've heard that term, the security, I hope you don't fly off - "to help make the world safer for everyone. Defenders can bring secure code review, threat modeling, patch validation, dependency risk analysis, detection, and remediation guidance into the everyday development loop so software becomes more resilient from the start. In the coming weeks" - which is part of what I thought was interesting. They're not quite ready yet. I think Mythos caught them a little flatfooted, and they're like, oh, wait, oh, yeah, we have something. What shall we call it?

In the coming weeks, we're working with our industry and government partners as we prepare to deploy increasingly more cyber-capable models as part of our approach to iterative deployment. That's right. So they're working on getting that thing together.

Okay. Nothing else they said on their introducing Daybreak page was surprising. But because they needed to say something, they did offer a couple bullet points, and then this talk about controlled containment. So they said: Focus. Focusing on the threats that matter. Prioritize high-impact issues and reduce hours of analysis to minutes with more efficient token usage. Okay; right. Patch safely, at scale: Generate and test patches directly in your repositories, with scoped access, monitoring, and review. And Verify. Verify every fix: Send the results and audit-ready evidence back to your systems to track and verify remediation. So this is all just boilerplate. This is what we've come to expect now. Look how quickly we got spoiled. This is what AI should do, if it's going to be taking care of that.

There was one final bit of interesting information. They said: Under "Choose the right level of access" and then "Contact the OpenAI team to align on the best model for your security workflows," they preview the three levels of access that they're talking about, saying GPT-5.5, which is the default level, which has standard safeguards for general-purpose use. Intended for general-purpose, developer, and knowledge work. Presumably that means anybody can have access to GPT-5.5.

Then you can go to Level 2, which is GPT-5.5 with Trusted Access for Cyber. They said: More precise safeguards for verified defensive work in authorized environments. Intended for most defensive security workflows, including secure code review, vulnerability triage, malware analysis, detection engineering, and patch validation. Okay. So you can do more with that one. So lowered or softened guardrails.

And then, finally, full strength at Level 3 is GPT-5.5-Cyber, where they says most permissive behavior for specialized authorized workflows, paired with stronger

verification and account-level controls. Intended for preview access for specialized workflows, including authorized red teaming, penetration testing, and controlled validation. So they're saying that in order for GPT-5.5 to be used by cyberthreat discovery, red teaming, penetration testing, and so forth, GPT-5.5 must be freed from its normal shackles which would otherwise prevent it from helping with such operations. Because an unshackled 5.5 could be abused by bad guys, the only model that can generally be used is the standard guardrailed 5.5 that apparently will resist some of the things you might ask it to do. If you want the guardrails dropped, then you need, they need to know why and who you are.

So that, you know, pretty much nothing at this point, I mean, we've got like a list of what we would like it to be doing is Daybreak, right, where the sky has yet to lighten because so far all we have is darkness. But we know what OpenAI is going to be doing. Basically it's a Mythos catch-up announcement, essentially. So, you know, they'll have something, too, is what they're telling us.

Something entirely different from Microsoft. I first picked up on this during last week's Windows Weekly, when Paul and Richard noted that Microsoft had been using an AI-driven system to uncover what they said on the podcast, mass quantities of bugs in Windows. And apparently not just any old run-of-the-mill random bugs, which we all know Microsoft fixes around a hundred or so of these days every month. Oh, no. These new bugs Microsoft was finding were what once would have been known as "showstoppers," so named because they would singlehandedly "stop the show" to prevent the release of software. So I thought, okay, that's interesting. And I mistakenly initially thought they were talking about Microsoft using Mythos. Nope. MDASH.

So having learned of this from Paul and Richard, I went searching and located Microsoft's posting from the previous day, which was last Tuesday, where Microsoft, for the first time, revealed that they have a, like, I guess I would call it a Super Mythos-like system of their own, only of course theirs is more better. The reveal was posted by Taesoo Kim, Microsoft's Vice President of Agentic Security. Okay, now, he's the real deal. In 2014, now Dr. Kim received his Ph.D. from MIT's EECS AI Research Lab. He's on leave currently from his professorship in the School of Cybersecurity and Privacy, and the School of Computer Science at Georgia Tech.

And it was he who led Team Atlanta, which took first place in DARPA's AI Cyber Challenge competition to build autonomous cyber reasoning systems to detect and remediate software vulnerabilities in open-source projects. I'm not going to enumerate his many awards. He's littered with them. Suffice to say that this looks like the guy that yes, indeed, you would like to get to build your autonomous vulnerability finding and reasoning system. And get him Microsoft did.

He posting last Tuesday was titled "Defense at AI speed: Microsoft's new multi-model agentic security system tops leading industry benchmark." And I'll say right off that it does start off with a bang. Dr. Kim writes: "Today Microsoft announced a major step forward in AI-powered cyber defense: Our new agentic security system helped researchers find 16 new vulnerabilities across the Windows" - get this - "networking and authentication stack, including four Critical remote code execution flaws in components such as the Windows kernel TCP/IP stack and the IKEv2 service." In other words, it doesn't get any more Internet facing than that. And these are critical RCE vulnerabilities in Windows TCP/IP stack.

So you might wonder when do we get that Windows update? Well, the answer is we got it the same day, during May's Patch Tuesday. So these things are fixed. They were - they weren't going to affect every Windows server on the planet, or you couldn't have talked about it then. They were in specific services that might not be used in every instance. So

we're probably okay. Four critical RCEs in the Windows kernel stack. So certainly better that Microsoft find these than somebody reverse engineering Windows networking.

So Kim continues, writing: "They used the new" - "they" meaning his team, the MSRC people. "They used the new Microsoft Security multi-model agentic scanning harness (codename MDASH) which was built by Microsoft's Autonomous Code Security team. Unlike single-model approaches, the harness orchestrates more than 100 specialized AI agents across an ensemble of frontier and distilled models to discover, debate, and prove exploitable bugs end-to-end.

"The results," he writes, "speak for themselves: 21 of 21 planted vulnerabilities" - and I'll explain what that is, it's actually an interesting test that they give to their human candidates - "found with zero false positives on a private test driver, that is, a software driver; 96% recall against five years of confirmed Microsoft Security Response Center (MSRC) cases in clfs.sys and 100% in tcpip.sys; and an industry-leading 88.45% score on the public CyberGym benchmark with 1,507 real-world vulnerabilities - the top score on the leaderboard, roughly five points ahead of the next entry."

He writes: "The strategic implication is clear: AI vulnerability discovery has crossed from research curiosity into production-grade defense at engineering scale, and the durable advantage lies in the agentic system around the model rather than any single model itself. Codename MDASH is being used by Microsoft security engineering teams and tested by a small set of customers as part of a limited private preview. This post explains how Codename MDASH works, what we shipped today, what we learned along the way, and how you can sign up for the private preview.

"The Microsoft Autonomous Code Security (ACS) team was assembled to take AI-powered vulnerability research from a research curiosity to production engineering at enterprise scale. Several members of this team came to Microsoft from Team Atlanta, the team that won the \$29.5 million DARPA AI Cyber Challenge by building an autonomous cyber-reasoning system that found and patched real bugs in complex open-source projects. The lessons learned from that work, especially the level of engineering required to make the frontier language models perform professional-level security auditing, are what our new multi-model agentic scanning harness (codename MDASH) is built around.

"Microsoft's code base is challenging for security auditing for a few reasons." And he has three bullet points. "First, massive proprietary surface. Windows, Hyper-V, Azure, and the device-driver and service ecosystems around them are private Microsoft codebases, not part of any commodity language model's training corpus, and are genuinely difficult to reason about. Kernel calling conventions, I/O Request Packets and lock invariants, Inter-Process Communication trust boundaries, and component-internal idioms do not yield to pattern matching. On this surface, a model must actually reason.

"Second point, DevSecOps at scale. Every finding has a real owner, a triage process, and a Patch Tuesday to land on. There is no quiet drawer for speculative findings. If a tool produces noise, the noise is everyone's problem.

"And finally, high-value targets. Windows, Hyper-V, Xbox, and Azure serve billions of users. The payoff for finding a single difficult bug is unusually high, and so is the cost of a false positive in a tier-one component."

He says: "The findings in this post are the result of a close collaboration between ACS, Microsoft Offensive Research and Security Engineering, and Microsoft Windows Attack Research and Protection." Those acronyms are MORSE and WARP. And he says: "MORSE and WARP own the deep, hard end of Windows offensive research; ACS brings the AI-powered discovery and validation pipeline. Together, the teams have collaborated to build a mature harness."

Okay. I now want to share what he explains about the structure of this startlingly complex agentic system which Microsoft has designed and assembled. This is going to sound more like science fiction actually than reality. A year ago it would have been regarded as a late April Fool's joke posting. Today I'd imagine that Microsoft's competitors are combing through it searching for hints. So get a load of this.

He writes: "A useful mental model is to think of it as a structured pipeline that takes a code base and emits validated, proven findings." Okay. Pipeline. Five stages. "Prepare stage: Ingests the source target, builds language-aware indices, and then draws the attack surface and threat models by analyzing the past commits. The scan stage: Runs specialized auditor agents over candidate code paths, emitting candidate findings with hypotheses and evidence.

"Third, the validation stage: Runs a second cohort of agents" - get this, the debaters - "that argue for and against each finding's reachability and exploitability. The fourth, de-dup stage: Collapses semantically equivalent findings (for example, patch-based groupings)." And finally, "The prove stage: Constructs and executes triggering inputs where the bug class admits it. The prove stage validates the pre-condition dynamically and formulates the bug-triggering inputs to prove existence of vulnerability."

And he says: "Three properties make this work in practice: An ensemble of diverse models that are effectively managed by codename MDASH. No single model is best at every stage. The multi-model agentic scanning harness runs a configurable panel of models. That includes state-of-the-art models as the heavy reasoner, distilled models as a cost-effective debater for high-volume passes, and a second separate state-of-the-art model as an independent counterpoint. Disagreement between models is itself a signal. When an auditor flags something as suspect and the debater can't refute it, that finding's posterior credibility goes up."

Then we have specialized agents. "An auditor does not reason like a debater, which does not reason like a prover. Each pipeline stage has its own role, prompt regime, tools, and stop criteria. We don't expect one prompt to do everything. We don't expect one agent to recognize, validate, and exploit a bug in a single pass. Codename MDASH has more than 100 specialized agents, constructed through deep research with past common vulnerabilities and exposures (CVEs) and their patches, working independently to discover the bugs, and their auditing results will be ensembled as a single report."

And then "End-to-end pipeline with extensible plugins. The pipeline is opinionated, but it is not closed. Plugins let domain experts inject context the foundation models cannot see on their own - kernel calling conventions, IRP rules, lock invariants, interprocess communication trust boundaries, codec state machines. The CLFS proving plugin we describe below is one such example: a domain plugin that knows how to construct a triggering log file given a candidate finding. For example, the Windows team extended reasoning with custom code analysis database, or CodeQL database, can be also leveraged.

"The payoff for this architecture is portability across model generations. The pipeline's targeting, validation, de-dup, and prove stages are model agnostic by construction, which allows the harness to get the best of what any model has to offer. When a new model lands, A/B testing it against the current panel is one configuration flip. When a model improves, the customer's prior investment - scope files, plugins, configurations, calibrations - all carry over, allowing customers to ride the frontier of security value."

Wow. Everyone knows that the last thing I am is a Microsoft apologist. I'm probably harder on them than I am on any other major player in our industry. One reason for that is that their behavior remains crucial to the functioning of much of the world. The other reason is that they're so big and so wealthy that it always seems that they should be able

to do a better job if they only cared to do so. I have no doubt that they are filled with very good people. But there's an institutional inertia that often doesn't appear to be producing the best outcomes for their customers.

But in this case, holy crap! If we believe all of this, they have really built something truly significant here, and there's more. Get this. They wrote: "To evaluate bug-finding capabilities of the multi-model agentic scanning harness you need to first ground on code that has never been seen by a model." Right? And we were talking about this just recently. Maybe one of the bugs that Mythos saw was actually it remembering something very similar. Not the same, but it may have contained it in its training.

He wrote: "This eliminates the possibility that a model 'learned the answers to the test,'" as he put it. "We scanned StorageDrive, a sample device driver used in Microsoft interviews of offensive security researchers. The driver contains 21 deliberately injected vulnerabilities, including kernel use-after-frees, integer handling issues, IOCTL validation gaps, and locking errors. Because StorageDrive is a private codebase that has never been published, we can safely assume it was not included in the training data of modern large language models.

"We ran the MDASH harness in its default configuration against StorageDrive. The results were striking: All 21 ground-truth vulnerabilities were correctly identified, with zero false positives. This simple test shows that the reasoning and vulnerability discovery capabilities of codename MDASH can approximate professional offensive researchers." And it doesn't get tired, and it can go 24/7/365.

"We then used the harness to conduct a security audit of the most security-critical part of Windows, namely, Windows' TCP/IP network stack. Right? I mean, that's what's hooked to the Internet. Across the Windows network stack and adjacent services, today's Patch Tuesday includes 16 CVEs our engineering teams found using codename MDASH. These vulnerabilities are 10 kernel-mode, six user-mode. The majority are reachable from a network position with no credentials."

Okay. The paper then takes the deep dive into two of the 16 vulnerabilities that were found and fixed. It provides way more detail than we need for the podcast, but the preface will give everyone a sense for what they are.

He just wrote: "The two findings below are characteristic of what the new Microsoft Security multi-model agentic scanning harness pipeline can do that a single model harness cannot. The first is a kernel race-condition use-after-free that requires reasoning about object lifetime across non-trivial control flow and three independent concurrent free paths. The second is an alias-aliasing double-free that spans six source files and is only visible against the contrast of a correctly handled site elsewhere in the same code base."

Okay. So stepping back from what gives all the appearance of being a significant achievement and an advancement, I mean, a bona fide advancement in automated vulnerability discovery at scale - and one that cannot come too soon, of course, as we know, for the Windows codebase. Since Windows source code is closed, we don't know objectively that OpenAI's Daydream, I mean Daybreak, or Anthropic's Mythos would not also have been able to find these problems. We don't know for sure. But Kim appears certain that no single model could do so. And this is his pedigree, so I'm inclined to trust that. Although obviously he has a pro-Microsoft bias. But this is also related to the approach that he took to win the DARPA prize.

And one of the beauties of this system that Microsoft has created is that it appears, as he said, to be model agnostic. We don't know whether Microsoft has their own internal models or much about them. But this assumes that they can use any model and plug it

into this. So it might well be, you know, using OpenAI's or Anthropic's models running as its agents.

In any event, I'm sure everyone understands why we needed to talk about this today. This is truly huge. I mean, imagine Patch Tuesday going away because there's nothing to patch. Instead of, oh, a hundred things this month and a hundred things last month. I've got no doubt that it's going to take Microsoft some time for what they appear to insist upon calling "Codename MDASH." You know, it's got to rummage around throughout their truly massive and buggy codebase. But once we emerge on the other side of that, Windows has at least the chance of leading the world in security rather than itself apologizing constantly for all of the problems that it has.

As Kim wrote: "AI vulnerability discovery has crossed from research curiosity into production-grade defense at enterprise scale." And given the evidence as presented, I see no trace of exaggeration there. It's going to be interesting when we get to the point where some future AI is able to say to Microsoft's security group: "Uh, guys? You realize that our Edge browser is needlessly leaving all of its users' login URLs, usernames and passwords decrypted in RAM for no reason; right?" You know, we're not there yet because that wasn't a bug. But really looking like AI is going to forever change the landscape of security of software, Leo.

Leo: Yeah, yeah.

Steve: And it just, I mean, and boy, this happened fast.

Leo: Yeah, it's amazing. What a world. Well, there you go. I'm sure this is not the last time we'll be talking about AI security tools. They're pretty amazing, and they're out there.

Steve: Wow.

Leo: Yeah.

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/>