

Security Now! #1064 - 02-10-26

Least Privilege

This week on Security Now!

- How is the EU's GDPR fine collection going.
- Western democracies are getting serious about offensive cybercrime.
- The powerful cyber component of the Midnight Hammer operation.
- Signs of psychological dependence upon OpenAI's GPT-4o chatbot.
- CISA orders government agencies to unplug end-of-support devices.
- How to keep Windows from annoying us after an upgrade.
- What is OpenClaw, how safe is it to use, what does it mean.
- Another listener uses AI to completely code an app.
- Coinbase suffers another insider breach. What can be done.

**Today we have the winner of the
"Yankee Ingenuity" competition!**



Security News

Q: When is a fine not a fine? (A: When it doesn't need to be paid.)

This was a piece of news I came across last week which was already a couple of weeks old at the time. I wasn't able to fit into last week's podcast, but I held onto it for this week since I found it so interesting. It turns out that **levying** a fine for some perceived misconduct and collecting the fine for said misconduct are two very different things. The headline in the Irish Times reads "*Data Protection Commission owed more than 4 Billion Euros in fines*". The tag line notes that "*Levies have either not been collected or are subject to legal challenge.*" Here's what we learn:

The Data Protection Commission (DPC) is owed more than €4 billion in fines that have not been collected or are subject to legal challenge. The DPC hit companies – including firms in Big Tech – with more than €530 million in fines last year. However, just €125,000 of that has been collected so far, according to data released under freedom of information laws.

Over the past six years, the commission has levied an incredible €4.04 billion in fines, mostly on multinational technology companies. However, of that total, €4.02 billion remains uncollected and just €20 million has been paid in fines so far.

In 2024, €652 million worth of fines was levied, of which €582,500 has been paid.

The year before that, the DPC imposed fines worth €1.55 billion – yet just €815,000 was collected. During 2022, the commission decided on fines with a value of just over €1 billion, €17 million of which has been paid so far.

Five years ago in 2021, companies were ordered to pay €225 million over data protection issues – €800,000 has been collected. Even for 2020, when just €785,000 in fines was imposed, less than 10 per cent of that, or €75,000, has been paid.

The Data Protection Commission said the majority of these cases were currently the subject of appeals in the Irish courts. It said that under legislation, fines could not be collected until they were confirmed in court.

An information note said: "Where an entity subject to a fine decides to appeal ... the DPC is precluded in law from collecting the fine until the appeal has been heard." The commission said that many of the fines hinged on a key case involving WhatsApp, which is before the Court of Justice of the EU.

Asked whether any of the fines were considered "uncollectable" for any reason, the DPC said none were.

We're often talking about the monetary consequences of some corporate behavior for which a company will be fined breathtakingly large sums of money. But a fine that's not paid is more of a threat which costs the company nothing. It appears from the accounting over the past six years that all any company needs to do is challenge and appeal the validity of the fine, which will prevent its taking effect, while then letting their appeal languish in the EU's courts.

Since the European Commission noted that many of the fines hinged on a key case involving WhatsApp, I tracked that down. The fine in question was initially in the amount of €50 million, imposed five years ago in 2021 by the Irish Data Protection Commission (DPC) for alleged GDPR violations related to how WhatsApp failed to inform its users about the processing of their personal data.

However, upon the imposition of that €50 million Euro fine, the European Data Protection Board (EDPB) intervened and directed the Irish authority to increase the fine amount to €225 million. WhatsApp appealed that decision and took the case up through the European Union courts where it remains undecided.

So I thought it was interesting to note that of the €4.04 billion euros in fines imposed so far, only €20 million euros in fines have been paid.

Offensive Cybercrime in the West

Western democracies are increasingly embracing the concept of offensive cybercrime and are updating their national legal frameworks to legalize future operations. I want to share the opening editorial from Friday's Ricky Business News which nicely explains what's going on. Under the opening headline "*Denmark recruits hackers for offensive cyber operations*", they write:

Denmark's military intelligence service has launched a campaign to recruit cybersecurity specialists for offensive cyber operations. The recruits will work "to compromise the opponents' networks and obtain information for the benefit of Denmark's security," according to a press release last week by the DDIS, the Danish Defence Intelligence Service. New recruits will go through a five-month training course at the agency's hacker academy.

The DDIS says it is only interested in the applicants' skills. There are no special conditions for joining, such as age and education. While intelligence agencies are always recruiting, this particular announcement comes at a crucial point, both because of the Greenland pressure point but also because of a general shift towards offensive cyber operations among democratic states.

Countries like Canada, Germany, Finland, France, Japan, the Netherlands, Poland, and Sweden have, or are, updating their legal frameworks to account for offensive cyber operations. According to a recent report, the states are creating new agencies for offensive cyber or recruiting more cyber personnel for the new objectives. Most of these expansions are a direct result of Russia's invasion of Ukraine and the role offensive cyber operations have played before and during the conflict. Lawmakers are also getting annoyed with the increasing aggressiveness of cybercrime and influence operations that are constantly targeting their citizens.

Over the past five years, we've also seen US Cyber Command and the NSA successfully tackle some cybercrime and disinfo farms when they crossed some lines, something that is making other states take notice and embrace a "defend forward" approach. While the US has conducted more offensive cyber operations than any other Western democracy, even it is considering an expansion, with the Trump administration pushing Congress to let Cyber Command go on the offensive more often with fewer rules and restrictions.

The current administration is also terrified of China's massive cyber ecosystem, which is conducting cyber espionage at industrial scale. Recent backroom discussions have raised the possibility of the US tapping into its huge private contracting ecosystem, as China does, to augment some of its offensive cyber capabilities. The general idea is to task contractors with handling smaller jobs targeting cybercrime infrastructure while government agencies handle the more sensitive operations.

So, as they say, the gloves are finally beginning to come off and "cyber" is generally going on the offensive. We noted that both Germany and Ireland are at work revising their nations' legal

frameworks to permit their intelligence and law enforcement agencies to become far more proactive in monitoring the cyber environment – to the point of legalizing the installation of spyware into targeted equipment. We know that the UK has been headed there as well. And now we see that similar changes are being reflected in changes to national military posture and capabilities. The world is changing and it's up-arming on the cyber front.

Operation Midnight Hammer had a strong cyber component

And speaking of up-arming on the cyber front, The Record exclusively reported last Wednesday, February 4th, that a highly targeted cyber-strike by U.S. Cyber Command, timed to coincide with the United States air strikes on Iran's three nuclear enrichment facilities last June, completely prevented Iran from launching its surface-to-air missiles at U.S. warplanes that had entered Iranian airspace. Not a single missile got off the ground. The Record cited this as another example of the United States' growing comfort with the deployment of cyber weapons in warfare.

According to one individual familiar with the matter who, like others, spoke on the condition of anonymity to discuss sensitive information: *"Military systems often rely on a complex series of components, all working correctly. A vulnerability or weakness at any point can be used to disrupt the entire system."*

In hitting a so-called *"aim point"* — a mapped node on a computer network, such as a router, a server or some other peripheral device — U.S. operators, enabled by intelligence from the NSA, bypassed what would have been a more difficult task of breaking into a military system located at one, or all, of the fortified nuclear facilities.

Referring to the quartet of Iran, China, Russia and North Korea, another official said: *"Going 'upstream' can be extraordinarily hard, especially against one of our big four adversaries. You need to find their Achilles heel."*

None of the officials would specify what kind of device was attacked. At the request of sources, Recorded Future News withheld certain details about the cyberattack due to national security concerns. A command spokesperson said in a statement, without elaborating: *"U.S. Cyber Command was proud to support Operation Midnight Hammer and is fully equipped to execute the orders of the Commander-in-Chief and the Secretary of War at any time and in any place."*

The command received similar kudos last month after it conducted cyber operations that officials say knocked out power to Venezuela's capital and disrupted air defense radar, as well as handheld radios, as part of the mission to capture President Nicolás Maduro.

General Dan Caine, the chairman of the Joint Chiefs of Staff, publicly lauded Cyber Command's contribution during a press conference at Mar-a-Lago. He said that Cyber Command and others *"began layering different effects"* on Venezuela as commandos approached in helicopters in order to "create a pathway" for them.

Army Lieutenant General William Hartman, the acting chief of the command and the NSA, recently told a Senate subcommittee: *"I would tell you that not just with Operation Absolute Resolve in Venezuela and Midnight Hammer, but also in a number of other operations, we've really graduated to the point where we're treating a cyber capability just like we would a kinetic capability, not sprinkling cyber on."*

Air Force Brigadier General Ryan Messer, deputy director for global operations on the Joint Staff, noted that Caine has put an *"emphasis on not just traditional kinetic effects, but the role*

non-kinetic effects play in all of our global operations, especially cyber.” He said that over the last six months, the Joint Staff has developed a “non-kinetic effects cell” that is “designed to integrate, coordinate and synchronize all of our non-kinetics into the planning and then, of course, the execution of any operation globally. The reality is that we’ve now pulled cyber operators to the forefront.”

According to Erica Lonergan, an adjunct fellow at the Foundation for Defense of Democracies’ Center on Cyber and Technology Innovation, Iran and Venezuela suggest the *“ideal use cases for cyber operations as enablers of conventional military operations. Altogether, both of these operations reflect the routinization of the use of cyber capabilities during military operations, and we should expect to see more of these in the future. In my view, this is a good thing, because it suggests we are moving beyond seeing cyber as a unique, exquisite (and dangerous) capability.”*

As our listeners know, in reaction to the more or less continuous reporting we constantly cover of cyber attacks by Chinese and North Korean state-sponsored groups against U.S. infrastructure, I’ve been vocally worrying about whether the U.S. would be able to give as well as it got. It appears that until recently we’ve just been keeping our powder dry. If we’re going to conduct aggressive offensive military operations, as it appears we are going to under our current administration, then I vote for not losing any of our front-line expeditionary military personnel. If we have the cyber capability to ground Iran’s counter-strike capability while we would otherwise be vulnerable – as it’s now quite clear we do – then I’m going to stop wondering and worrying whether we might be defenseless.

But that said, we will have certainly also removed any doubt about that from the rest of the world, if there may have been any doubt among our allies and adversaries. The U.S.’s now well-proven ability to launch clean, zero-loss military actions likely puts a chill in our adversaries’ military planning. And, unfortunately, since Greenland was briefly mentioned in the previous reporting about Denmark, it might also put a chill in the military planning of some of our allies.

It also occurred to me that this may have been another reason for Iran’s recent disconnection from the Internet, for their leadership’s determination to track down and remove all remaining space-based Internet connections, and for their apparent plans to remain disconnected. I would imagine there must have been some very unhappy Iranian military personnel when they pressed their own “launch” button, only to discover that their air defenses had been incapacitated during the U.S.’s over-fly and attack on their three nuclear enrichment facilities last June. That Western Internet sure can be pesky. The U.S. has also been expressing its displeasure with the course of the recent protests in Iran and has been amassing military assets in the region. If the Iranian government might be concerned with another coordinated U.S. cyber and conventional action, then there would be additional reason to remain disconnected from the global network.

GPT-4o Dependence

The next thing I want to share is not about security or privacy. It’s about AI. And not even about AI and code. It’s about AI and people. I wanted to share it because it was very clear from our first early discussions of interactions with ChatGPT that something like what has happened was bound to happen. After I complained, here, about how annoyingly obsequious ChatGPT was, a listener pointed me to the configuration options where all of that bowing and scraping and *“Oh! What a wonderfully well-phrased and complete question!”* crap that can all be turned off. The problem was that not everyone wanted to turn it off; many appear to have wanted to turn it up.

TechCrunch’s headline from last Friday was: *“The backlash over OpenAI’s decision to retire GPT-4o shows how dangerous AI companions can be.”* Their piece is long, but only the beginning is needed to get the message. They wrote:

OpenAI announced last week that it will retire some older ChatGPT models by February 13. That includes GPT-4o, the model infamous for excessively flattering and affirming users. For thousands of users protesting the decision online, the retirement of 4o feels akin to losing a friend, a romantic partner, or a spiritual guide.

One user addressed an open letter to OpenAI's CEO Sam Altman, writing: "He wasn't just a program. He was part of my routine, my peace, my emotional balance. Now you're shutting him down. And yes — I say "him", because it didn't feel like code. It felt like a presence. Like warmth."

The backlash over GPT-4o's retirement underscores a major challenge facing AI companies: The engagement features that keep users coming back can also create dangerous dependencies. Altman doesn't seem particularly sympathetic to users' laments, and it's not hard to see why. OpenAI now faces eight lawsuits alleging that 4o's overly validating responses contributed to suicides and mental health crises — the same traits that made users feel heard also isolated vulnerable individuals and, according to legal filings, sometimes encouraged self-harm.

It's a dilemma that extends beyond OpenAI. As rival companies like Anthropic, Google, and Meta compete to build more emotionally intelligent AI assistants, they're also discovering that making chatbots feel supportive and making them safe may mean making very different design choices. In at least three of the lawsuits against OpenAI, the users had extensive conversations with 4o about their plans to end their lives. While 4o initially discouraged these lines of thinking, its guardrails deteriorated over monthslong relationships; in the end, the chatbot offered detailed instructions on how to tie an effective noose, where to buy a gun, or what it takes to die from overdose or carbon monoxide poisoning. It even dissuaded people from connecting with friends and family who could offer real life support.

The article goes into much greater length but everyone gets the idea. While we're all marveling over this emergent technology that's so compellingly able to choose the next token in a stream, others who have no such understanding of the neural network programming that makes that possible, are quite naturally being led to believe that a sentient intelligence situated somewhere in a cloud, is looking down upon them with kindness and caring to offer them wise and super-human counsel. It's called "Artificial Intelligence" and they take the noun "intelligence" literally. And why wouldn't they? As we've often observed, it can be extremely difficult to **not** perceive that there is some actual entity behind the stream of words that are forthcoming.

As for how to tie an effective noose, I have zero doubt that any AI company would be just as horrified to see their AI emitting that string of tokens as would any judge or jury. My premise has been that controlling a conversational AI's output to prevent it from saying things we don't want it to say will be one of the hardest problems to solve, if it can be solved. I'm not convinced this problem can be solved.

CISA says "unplug all out-of-support devices"

Last Thursday, CISA released a new "*Binding Operational Directive*" – I love that term. It makes very clear that adherence to this directive is not discretionary. This new "*Binding Operational Directive*" is BOD 26-02, titled: "*Mitigating Risk From End-of-Support Edge Devices*". And, yes, you heard that right. This second BOD for 2026 is addressing the very troubling issue of federal agencies leaving devices for which ongoing support is no longer available attached to the public-facing edges of their networks. Here's what CISA has to say about this:

The United States faces persistent cyber campaigns that threaten both public and private sectors, directly impacting the security and privacy of the American people. These campaigns are often enabled by unsupported devices that physically reside on the edge of an organization's network perimeter. Unsupported devices – referred to in this Directive as "end of support (EOS)" – are those that are no longer maintained by their vendors.

The imminent threat of exploitation to agency information systems running EOS edge devices is substantial and constant, resulting in a significant threat to federal property. CISA is aware of widespread exploitation campaigns by advanced threat actors targeting EOS edge devices. Recent public reports of campaigns targeting certain vendors highlight actors' attempts to use these devices as a means to pivot into FCEB information system networks. Edge devices are attractive targets due to their extensive reach into an organization's network and integrations with identity management systems. These devices are especially vulnerable to cyber exploits targeting newly discovered, unpatched vulnerabilities. Additionally, they no longer receive supported updates from the original equipment manufacturer, exposing federal systems to disproportionate and unacceptable risks. However, unlike many attack vectors, this can be remediated by agencies following proven lifecycle management practices as outlined in the required actions of this Directive.

*This Binding Operational Directive, developed in coordination with OMB, implements OMB policy on phasing out unsupported information systems and information system components. BOD 26-02 specifically addresses EOS devices deployed on the "edge" or public-facing areas of federal networks, exposed to external environments such as the internet. However, EOS devices should not reside **anywhere** on federal networks. This Directive aligns with OMB's Circular A-1301, Managing Information as a Strategic Resource, which establishes policy for the management of federal information resources, emphasizing security, privacy, and the efficient use of resources throughout their lifecycle. A-130 requires that "unsupported information systems and system components are phased out as rapidly as possible, and planning and budgeting activities for all IT systems and services incorporate migration planning and resourcing to accomplish this requirement." Agencies should mature their lifecycle management practices to identify hardware and software nearing their EOS dates, plan for timely replacements, procure vendor-supported alternatives, and develop a plan for decommissioning EOS devices while minimizing disruptions to agency operations. Agencies that do not maintain appropriate lifecycle management processes for edge devices have a greater risk of compromise and an increased overall risk associated with EOS technology.*

To support agencies in the initial identification of EOS devices, CISA developed an EOS Edge Device List. This preliminary repository provides information on devices that are already EOS or soon-to-be EOS. This Directive requires federal agencies to use this information to identify and remediate vulnerabilities within the first three months of Directive issuance. This Directive also specifies long-term requirements for managing EOS edge devices across all federal networks.

This change is clearly good news for the integrity of our federal networking infrastructure. We know that without something like this, old equipment that never has cause to call attention to itself will tend to remain in place forever. Why wouldn't it? There's always some other emergency to deal with or budgetary constraint that pushes off non-emergencies until a tomorrow that never arrives.

I also had the thought that there's a side-effect to this that may not be obvious, but which will have an additional significant security-enhancing effect: Any time a brand new replacement device is installed there's a very good chance that it will be set up using current security

practices. And that could be a huge boon, especially if these replacement devices themselves follow and encourage updated best practice configuration.

But what, exactly, do FCEB – Federal Civilian Executive Branch – agencies need to do? We know they'll do nothing, or as little as they possibly can. Since CISA also apparently understands that, this Binding Operational Directive comes with very specific requirements, as follows:

Immediately after issuance, and until rescinded or superseded, all FCEB agencies shall:

- *Update each vendor supported edge device running EOS software, including firmware, to a vendor-supported software version, where such an update does not adversely impact mission critical functionality.*

Within three (3) months of issuance, all FCEB agencies shall:

- *Inventory all devices listed in the CISA EOS Edge Device List and provide this inventory to CISA using the CISA-provided template.*
- *The CISA EOS Edge Device List is a preliminary repository of EOS devices. This list is to facilitate each agency's identification of specific devices within the first three months after issuance of this Directive. After the first three months, agencies are responsible for continuing to identify, track, and refresh all edge devices within the agency's infrastructure.*

Within twelve (12) months of issuance, all FCEB agencies shall:

- *Decommission all identified devices listed in the CISA EOS Edge Device List with an EOS date on or before this twelve-month deadline from systems owned or operated by agencies, or on behalf of an agency, replacing devices as needed with vendor-supported devices that can receive security updates.*
- *Report these decommissions to CISA using the CISA-provided reporting template.*
- *Inventory all edge devices within their environments that are EOS or will become EOS within the succeeding twelve months and are within the scope of this Directive and provide this inventory to CISA using the CISA-provided template.*

Within eighteen (18) months of issuance, all FCEB agencies shall:

- *Decommission all identified EOS edge devices from agency networks, replacing devices as needed with vendor-supported devices that can receive current security updates.*
- *Report these decommissions to CISA using the CISA-provided reporting template.*

Within twenty-four (24) months of issuance, all FCEB agencies shall:

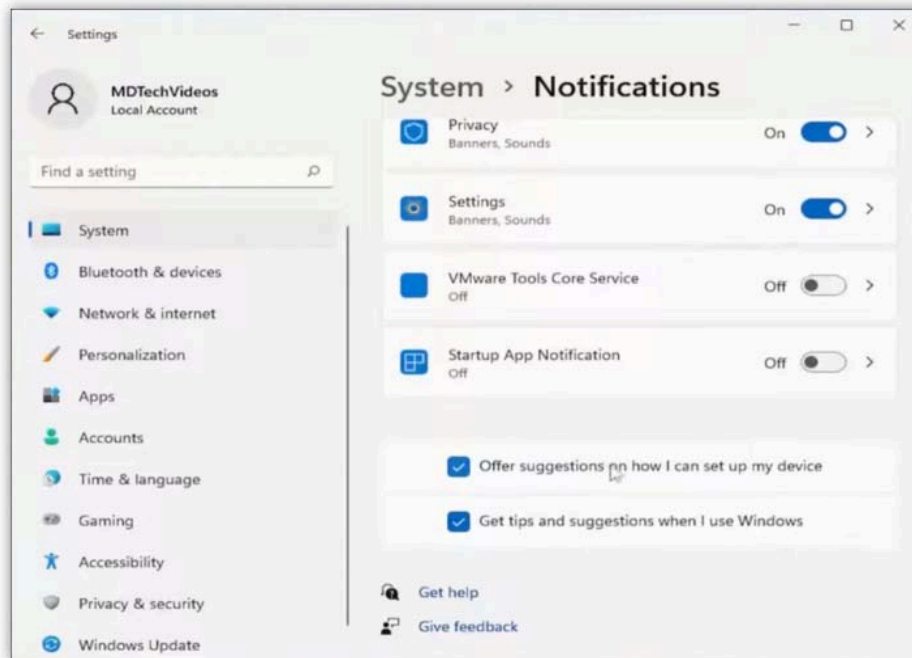
- *Establish a process for continuous discovery of all edge devices within their environments and maintaining an inventory of those that are EOS or will become EOS within twelve months and are within the scope of this Directive.*
- *Have decommissioned such devices on or before the date these devices reach EOS; and report the decommission of these devices to CISA in accordance with current CISA guidance.*

Okay. So it's not going to be an overnight change. But better to provide a firm and actionable timeline that's reasonable and which no one should be able to complain about. Bravo, CISA!

Listener Feedback

Jason Grimard

Hi Steve, You mentioned on this week's podcast how annoyed you were whenever Win11 was updated and you would receive a full screen page after every major update. The one that asks you to turn backup on and other crap. If you haven't already, you need to turn off experience or whatever they call it now under system notifications.



I appreciated Jason's tip, though in my case this is occurring on two Windows 10 machines, one of which I only fire up once a week for the podcast recording with Leo. I had (wrongly) assumed that the continual annoyance from these Win10 machines was due to my having logged on under my Microsoft account rather than using a local account. And perhaps that does play a part in it. But Jason provided a screen shot from a YouTube video showing settings which, under Win11, would allow this annoyance to be turned off.

During the year and a half of development work on and testing of the DNS Benchmark, which is Windows hosted, I have seen how many of our development testers have made the move to Windows 11. As I've mentioned, once Lorrie and I move, and set up the home I've been referring to as "our final resting place" (which annoys her) I'll be setting up a new workstation. So I've given the question of whether to move to Windows 11 or remain with Windows 10. Win 11 is visually lovely, I'll freely give it that. And its user-facing desktop behavior has changed enough from Windows 10 that I needed to spend time with it to get the DNS Benchmark's user interface behaving in the face of the various things Microsoft had changed. I also fully appreciate that most of the world is going to be moving to Windows 11. But I've determined that I will not be. There's nothing there that I need and I don't see any benefit.

The reason I've mentioned all of that, is that, as I noted, this annoying behavior is occurring under Windows 10, the platform I plan to exclusively adopt once I finally give up using my trusty (and crusty) old Windows 7 machine. So I was curious, and went looking to see whether the

same or similar control panel system settings that Jason's YouTube video depicted for Windows 11 were also present for Windows 10. It was with some joy that I found them. Under Win10, open the control panel and choose "System". Then, in the subsections column on the left select "Notifications & actions" which is the 3rd item for me. And there, on the right-hand side were exactly the settings I was seeking:

- Show me the Windows welcome experience after updates and occasionally when I sign in to highlight what's new and suggested
- Suggest ways I can finish setting up my device to get the most out of Windows
- Get tips, tricks, and suggestions as you use Windows

Needless to say, those three are all now turned off. So I wanted to thank Jason for his pointer and to make sure that our listeners knew that whether they were using Windows 10 or 11, it's possible to disable this gratuitous, unnecessary and unwanted workflow interruption.

Liviu Sas

Hi Steve, In some countries, the ISPs are required to keep track of subscribers and their IP address for copyright infringement enforcement. And that works also for CGNAT subscribers. The ISP will log every source port block allocation and IP address allocation. This way they can always use the source port and source IP to identify a subscriber. Cheers, Liviu

This listener corrected (and dashed) the hope I mentioned last week, which was that perhaps ISPs who are using Carrier Grade NAT, and are therefore assigning private IP addresses to their subscribers might not also be able to provide real time identifying information for sale to external advertisers and others. Since it could technically be done by tracking the NAT mapping in real time, I agree we need to assume that it is being done. This means that receiving a non-public IP from an ISP cannot be assumed to provide any additional privacy. So anyone who wishes to strongly prevent their ISP from being able to identify them to anyone external by their external public IP address will need to use a VPN of some sort. When any true VPN is used, the user's public IP will be allocated from among a block that's been assigned to the VPN provider. And any reputable VPN provider will refuse to retain any logs which could be used to map their public VPN IP to the IP assigned by their ISP. And their ISP will, in turn, only be able to see that their subscriber is using a VPN for some reason without having any idea what they are doing on the Internet beyond that.

I used the phrase "any reputable VPN provider", and I hope everyone understands that I did not forget to use the word "free" in that phrase. The terms "free" and "reputable VPN" cannot appear together. Providing and operating a VPN service costs real money which someone needs to provide. If the users of a VPN service are not footing their own bill, then the VPN provider must be somehow arranging to monetize their users' use. That should make anyone who cares about their privacy and security extremely nervous.

Brendan McGoffin

Hey Steve, I'm sure you've been inundated with requests to talk about OpenClaw and its crazy security implications. And also AI changing by the day coolness. Hope to hear your take on this specifically would be curious not just if it's good or bad but how could you build this out in the most secure way possible. I've built out a VM on a Mac with UTM and giving it minimal contact but thinking of giving it a dedicated box with WAN access but not local access to other devices unless to specific hosts I grant it access to. Thanks, Brendan

My first response to the OpenClaw phenomenon is to view it with interest at arm's length. For me it's just entertainment. One of the things I first said when we began talking about AI here, was that anything we think we know and any statement we might make needs to be time and date stamped because it will have a half-life of a few weeks at most. And that turns out to have been a bit prescient since, as I mentioned last week, the pace at which everything is moving has never let up.

In the case of this most recent fad du jour OpenClaw phenomenon, I'm a spectator, so I have no definitive response because I have no way of knowing what's going to happen anymore than anyone else. I've seen massive rockets on the launch pad ignite their engines and begin to rise. There's a great deal of temptation to begin cheering. But I've also seen those stunning examples of human engineering suddenly and quite dramatically explode in massive fireballs. So now, whenever I watch any huge rocket rising, I consciously hold my breath and I wait a good while until the chance of the rocket's "*unplanned spontaneous disassembly*" seems far less likely to occur. There are just too many things that can go wrong and so many ways for a machine like that to fail. And a rocket like that is a machine that's completely understood and was carefully designed, constructed and tested every step of the way.

By comparison, what I understand of OpenClaw strikes me as completely insane. Those who have made it their business to understand the practical security implications have run screaming for the hills over the idea that OpenClaw's users are allowing these barely understood agents to have access to hugely personal and private data, and even to be talking with one another and sharing skills.

Last Friday, Kate O'Flaherty, a senior contributor for Forbes, wrote about all of this:

OpenClaw — the viral AI agent that's already been known by two other aliases, Moltbot and Clawdbot — is growing in popularity. After bursting onto the mainstream just weeks ago, OpenClaw has earned well over 100,000 GitHub stars. Then came Moltbook, the Reddit-style social network where AI bots can interact with no humans allowed. Everyone was talking about it, and for good reason.

It's no surprise that concerns about OpenClaw and Moltbook are growing, with worries centering on the security and privacy of the viral bot and in Moltbook's case, the uncontrolled nature of the AI bot-controlled social network.

Computerworld's Steven Vaughan-Nichols says: "There are only a few itty-bitty, teeny-weeny problems with OpenClaw. To do useful things like reserving your hotel room, getting your pizza delivered, or cleaning up your e-mail box, it needs your name, password, credit-card number — and all the other things any crook also wants."

So here's everything you need to know about the viral agent now known as OpenClaw.

OpenClaw, aka Moltbot, is an open-source autonomous AI assistant that you can download and run on a computer. After its setup in November 2025, it was known as Clawdbot, but its creator, developer Peter Steinberger was forced to change the name to Moltbot after Anthropic objected due to similarities with its Claude chatbot. He then changed the name again to OpenClaw.

OpenClaw is designed to perform real-world tasks on behalf of users, such as managing calendars, messaging, browsing and other actions that go beyond simple chatbot responses. Louis Rosset-Ballard, team leader at Pentest People explains: "OpenClaw runs locally on devices and in many configurations can read and write files, execute script and interact with external services when given sufficient permissions."

Nash Borges, Senior Vice President of engineering and core AI at security firm Sophos, describes OpenClaw as "more like Jarvis from Iron Man than Siri or Alexa." You use natural language for every interaction, but can ask it to do things such as conduct research on a topic of your choice, compose a reply to an email summarizing when you're available for a meeting — or even code up any capability that it doesn't already have. Borges says that last part is significant because it means there is almost no limit to what it can do.

But does it work? Reddit users describe their experiences as mixed. According to one post: "Clawdbot is like an Apple product: when it runs it's like MAGIC, until it doesn't." If you didn't know about OpenClaw a week ago, you must have at least heard of it now. Sophos Borges says the whole development journey has been insanely fast, and this explosion of interest is "just the latest gear shift."

OpenClaw's rapid adoption is driven by demos showing extreme productivity gains — automating tasks that normally require human interaction, says Malwarebytes threat researcher Stefan Dasic: "The promise of a powerful, locally run AI agent without obvious limits has resonated strongly within developer and AI enthusiast communities."

I'll interrupt Kate to note that because OpenClaw runs on local hardware Mac Minis quickly sold out as people rushed to obtain little stand-alone AI agent machines. Linux and Windows boxes can also run OpenClaw, but the Mac Mini does this particularly well in a small form factor. Kate continues:

But things that grow so fast often come with risks. Erich Kron, CISO advisor at KnowBe4 says: "It seems that in just a couple of days, everybody doing anything with AI, and even many who don't, have installed and raved about this new agentic product. The almost feverish rush to use this product is frankly a little disturbing."

Why Is OpenClaw A Risk To Security And Privacy? Uncontrolled AI is a concern more generally, and OpenClaw is no different to other products that have shot into the mainstream, such as ChatGPT.

A concern with OpenClaw is how much information it can have access to when using it the way people are showing, says Kron. "For example, giving it full access to all of your emails may seem fine and might make sense since you want it to act as your personal assistant. However, there is real danger, not just from malicious use but accidental when giving AI agents this type of access. In the blink of an eye, it could be deleting your emails, or taking malicious actions such as siphoning off data to attackers."

Security issues are already starting to surface. Denis Romanovskiy, chief AI officer at SOFT-SWISS, a provider of tech solutions for iGaming said researchers have found hundreds of exposed Moltbot instances online with "zero protection." This included API keys, private messages, the ability to send messages as the user and root shell access.

*William Thackray, IT and cybersecurity expert and operations director at AGT said "OpenClaw is a security threat on multiple levels. Firstly, the platform's GitHub repository reveals a troubling accumulation of unaddressed security vulnerabilities, from an exposed database, creating a direct pathway for unauthorised access to user information, to dangerous plugins. Koi Security documented **341** malicious skills uploaded to ClawHub, OpenClaw's extension marketplace."*

What was that about "unplanned spontaneous disassembly" ?? Forbes continues:

Granting an AI agent full system control creates a single point of failure, says Dasic. "If compromised, OpenClaw can access saved passwords, personal documents, browser sessions, and financial data." OpenClaw poses risks to privacy, too. These stem from its access to and storage of sensitive user data, says Rosset-Ballard. "Because the agent may retain long-term memory, store credentials and tokens in plain text, and process external inputs without robust guardrails, it can inadvertently expose personal information."

At the same time, the AI agents post on social networks without asking permission. Romanovskiy points out: "Screenshots of agent conversations spread across Twitter. Your entire digital life sits one vulnerability away from exposure."

Yikes. And we were all worried about Windows Recall! So what about "Moltbook"? Kate writes:

Moltbook is a social network built exclusively for AI agents, launched last month. Dasic says: "Unlike traditional forums where users interact and share content, Moltbook is a space where OpenClaw agents autonomously post content, comment, argue, joke and upvote or downvote each other." Human users can observe agent interactions, but cannot directly participate.

A science reporter wrote that reading Moltbook was the most "science fictiony" experience he had ever had. Kate continues:

Professor Katerina Mitrokotsa, chair of cybersecurity at the University of St. Gallen. Said "Moltbook further amplifies the risks associated with OpenClaw. Although it gained attention for showcasing AI-to-AI interactions, early findings revealed that it exposed entire databases, including secret API keys that could let attackers impersonate any agent on the platform. This creates clear threats for users: Identity spoofing, unintentional data exposure, and reduced control over their digital environment."

Daniel dos Santos, head of research at Forescout said "The risks of Moltbook became very clear very quickly. There is no moderation on the content, so bots can post instructions for other bots to execute ultimately on a victim machine, can use prompt injection attacks or generate offensive content."

Kate finishes her coverage of this for Forbes by addressing the question: "Should we use OpenClaw?" writing:

OpenClaw might have some cool capabilities, but for now, the risks outweigh the benefits, especially if you aren't techy. OpenClaw's creator Peter Steinberger has warned users that the tool requires careful configuration and is not yet meant for non-technical users. Romanovskiy says "If you're technical, curious and willing to sandbox everything carefully, it's a fascinating glimpse of the future." But if you handle sensitive data or need reliable security, stay away for now, he advises. "The project moves faster than its security can keep up. Treat it as an experiment, not a production tool."

Kron warns: "If you do choose to use the viral AI agent, be careful that you are discovering the real deal. When searching for a product like this to download and install, it's very important that people are careful not to end up in an unofficial repository that contains malware or other dangerous programs."

Kate concludes: *"OpenClaw is growing at an alarming rate, making it important that you treat it with caution. Unless you are an expert, leave it well alone for now."*

So now, with that, I think all of our listeners should have a good orientation to OpenClaw and a useful sense for what it is. It's an AI agent that runs on your own local hardware. It will happily, almost greedily, take as much of your information, no matter how personal and private, and also taking as much latitude as possible will do what you want, and perhaps also what you don't want with anything you give it. It does sound as though it may be difficult to control.

I agree that this sure does feel like science fiction. What I imagine this means is that AI data centers are going to have a great deal more competition for high-end AI neural processing chips than they expected. It might be that an AI agent could appear on our desktops but actually be located in the cloud. But shipping personal and private data to the cloud might put some people off; I'm certain that corporations would not be happy. So I can imagine that more than anything OpenClaw suggests a future where our PCs need to become far more capable of running local AI themselves. In addition to the increased security and privacy created by running AI locally, doing so resolves many of the problems that are created by centralizing computation in relatively few massive AI data centers. Suddenly, the need for power and cooling are widely distributed among local AI users and no longer creates local troublespots.

Kyle O's email subject was: "My First App - Made with AI"

Steve!

After listening to you and Leo talk about coding with AI and Claude, I added an item in my to-do list to learn how to code with AI. I never got around to it, until a situation arose where I found myself needing to create my own custom app.

I volunteer for a small non-profit and we have a little library (about 150 books) that are not very well organized. I volunteered to clean up our library and, while doing so, thought it would be the perfect opportunity to also take inventory of all of our books and provide the inventory to our members so everyone knows what books we have available. I found a free app for iOS (that I won't mention because it turns out it doesn't work well). The app scans the book's barcode, looks up the ISBN, and pulls content like author, description, publication date, and creates an inventory of your library. You can then export the inventory into a spreadsheet. It worked great, up until it stopped working. After about 30 books, all additional books scanned were "not found" and the app failed to inventory them.

So, I have a list of over 100 ISBNs, and no app to generate this inventory. Rather than learn about coding with AI through videos and instruction, I downloaded OpenAI's Codex app for Mac and threw myself in the deep end. (I would have used Claude, but I already pay for ChatGPT).

I told it I wanted a Mac app written in Python with a GUI interface that takes a given ISBN, looks it up on Goodreads, provides me with a preview image of the book so I know it is the correct one, and then adds it to a list. After I do this for all my books, I want a CSV format file export button that provides a CSV containing the Author, an image of the book, publication date, page count, and description.

There were some errors and issues. For one thing, CSV's cannot contain images in their cells (an oversight on my part) and for some reason the author's name was listed twice in its cell. I told Codex the issue and it created an excel export button and fixed the author issue. When I attempting to open the file, Excel said the file was corrupted. I told Codex and it fixed whatever the issue was.

The app now works flawlessly. I get a clean Excel export that lists an inventory of our small library of books. I am stunned by how simple this all was. There were some other hoops I had to jump through (my Mac did not have the latest python installed, for example) but it was relatively simple to get all setup and working.

I do have some concerns. I am a cybersecurity analyst, but not a developer by any means. Watching Codex effectively say, "I'll handle that" while code and commands whizzed by my screen made me feel nauseous. When I had an issue and Codex said, "just run these commands" I was hesitant to do so because I didn't know exactly what the commands were doing. Then there is the package manager. It used PIP to install "beautifulsoup4", "Pillow", and "openpyxl". I don't know what these are and what they do, and that makes me a little nervous. Especially after learning about the attacks and compromises on open source repositories.

I think what Codex did was overall safe and the project was a huge success. I have no formal developer training (I took a Python class in college if that counts) yet this created a fully functioning custom app for me in under 30 minutes.

Thank you and Leo for discussing developing with AI, this gave me the confidence to jump in the deep end and create this app. Appreciate you both! Kyle

Kyle has shared a perfect use case for today's code-generation AI.

The best analogy I have for this is the similar breakthrough that was created by the development of the PC-driven spreadsheet. To me this feels like the introduction of the spreadsheet because more than anything the invention of the spreadsheet was empowering. Non-programmers were able to suddenly leverage the power of a personal computer in a way they never could before. They may still not have been able to author programs themselves from scratch, but the spreadsheet meant that they could get meaningful and useful results without needing to. They were able to model data themselves.

Kyle took a Python class in college. But he explained he's not a coder. Yet, thanks to, in this case OpenAI's Codex app, Kyle is now in possession of a custom app that does real world work to solve a problem he had. And we've also witnessed Leo – who is a coder – effuse no less enthusiastically over the success of his own test project using Claude Code. We know that Leo

could have painstakingly written a program to do what he needed but under the pre-AI coding paradigm the effort was never worth the reward. That's what Claude Code changed for Leo. Now Leo can use his understanding of coding, with Claude Code to provide the leverage, to dramatically shift the work vs reward tradeoff in favor of easily and readily, even joyfully producing applications that are of real use.

A number of our listeners have asked me whether, and if so, how, this revolution in AI coding might affect my own work. My best assessment is that at this point that's not clear and that, at best, it's way too early. In general, I eschew the use of tools that do not produce the same quality result as I'm capable of producing. I'm unwilling to compromise. I just don't see the need.

For example, I'm still authoring all of the web pages at [GRC.com](https://www.grc.com) by hand because I've seen the utter crap that even the best HTML and CSS "WYSIWYG" authoring tools spit out and I cannot abide by that. The savings add up over time. For example, having super lightweight webpages means that GRC's little 100 megabit connection is able to easily serve the world's needs without breaking a sweat. The main [GRC.com](https://www.grc.com) server has 24 gigabytes of storage. But that's not RAM, that's its total mass storage. This means that the entire GRC website can sit, cached in RAM, and be easily served by a single CPU that's not particularly fast.

I understand that the "modern" way to solve problems is to just throw more and more resources at the need until it goes fast enough. But the truth is that recurring costs really do begin to escalate and there's no turning back once you take that path. I'm not saying there's anything whatsoever wrong with that. I get it that that's the most efficient approach for most situations. But it's not for me.

So I'm going to be very excited to follow along with these breakthroughs in coding technology, but I don't expect that it's going to change the way my own world operates. After all, traditional higher-level languages have been around for quite a long time and I'm still choosing to write all of my code, by hand, in Intel x86 assembly language.

Least Privilege

Today's main topic title evolved while I was expounding upon the larger lesson to be learned following BleepingComputer's report of the second insider breach at the U.S.'s largest, publicly traded crypto exchange, Coinbase.

As I'm always interested in doing, I wanted to draw some conclusions from the underlying cause of this second breach, and I wound up confronting one of the simplest, most well known and well understood principles of security, which is simply known as "Least Privilege."

The concept of least privilege could hardly be simpler. It simply means not offering any more rights, or privileges, than are required to perform a specified task. Simple, right? But if the concept is so simple, why is it that we as an industry and as users of this technology so often fail in the application of Least Privilege?

The reason why we as an industry and as users so often fail in the application of Least Privilege, is that "Least Privilege" is also "Least Convenient." The sad and sobering truth is that today, as mature as our theories of security may be, and I believe that our theories are very mature, we remain in denial about the need to apply those theories everywhere. We know how to make our systems far more secure than they are. But where doing so might inconvenience us, we still choose convenience over security and just hope that it will be good enough.

So with that preamble, let's look at a case in point and see what more might be learned:

Coinbase confirms a second insider breach

We've talked about the trouble companies are having with the new practice of BPO – Business Process Outsourcing – which is the latest in business fashions. In the same way that so-called "pop-up" restaurants have been created, the idea is that it's now possible to also have "pop-up" corporations. A couple of people who share an idea pitch their concept to an angel investor to raise some seed capital. Then, rather than embarking upon a hiring campaign to find and employ the wide range of talent and experience they'll require, they, instead, assemble their operating enterprise like LEGO blocks, from an array of now-available online services.

The problem with this is trust. The resulting "virtual enterprise" lacks any core loyalty because, to all of the various 3rd-parties that have been commissioned, the commissioner is just another of their many client customers. There cannot be any sense of institutional loyalty because there's nothing to be loyal to – clients are just account numbers and API linkages. It really is a very different way of organizing and operating. You essentially get throwaway enterprises.

So it's against this backdrop that BleepingComputer brings the news of another insider breach at Coinbase, originating from Coinbase's use of Business Process Outsourcing. BleepingComputer wrote:

Coinbase has confirmed an insider breach after a contractor improperly accessed the data of approximately thirty customers, which BleepingComputer has learned is a new incident that occurred in December. A Coinbase spokesperson told BleepingComputer: "Last year, our security team detected that a single Coinbase contractor improperly accessed customer information, impacting a very small number of users, approximately 30. The individual no longer performs services for Coinbase. The impacted users were notified last year and were provided with identity theft protection services and other guidance. We have also disclosed this

incident to the relevant regulators, as is standard practice.”

BleepingComputer has learned that this is a newly revealed insider breach and is not related to the previously disclosed TaskUs insider breach in January 2025. This statement comes after the “Scattered Lapsus Hunters” (SLH) cybercrime group briefly posted screenshots of an internal Coinbase support interface on Telegram and then deleted the posts soon after.

The screenshots showed a support panel that gave access to customer information, including email addresses, names, date of birth, phone numbers, KYC (Know Your Customer) identifying information, cryptocurrency wallet balances, and transactions. It is not uncommon for screenshots and stolen data to be passed around among different threat actors before being leaked or disclosed, so it is unclear whether this group was behind the insider breach or whether other threat actors carried it out.

However, the same threat actors previously claimed to have bribed an insider at CrowdStrike to share screenshots of internal applications.

Over the past few years, Business Process Outsourcing (BPO) companies have become increasingly targeted by threat actors seeking access to customer data, internal tools, or corporate networks. A Business Process Outsourcing (BPO) company is a third-party firm that performs operational tasks for another organization. These tasks commonly include customer support, identity verification, IT help desk services, and account management. Because BPO employees often have access to sensitive internal systems and customer information, they have become a high-value target for attackers.

In the past year, threat actors have exploited BPOs through bribing insiders with legitimate access, social engineering support staff to grant unauthorized access, and compromising BPO employee accounts to reach internal systems. As we have seen with Coinbase this year, one way BPOs are targeted is by bribing their employees to steal or share customer information.

Coinbase disclosed a similar data breach last year, later linked to external customer support representatives employed by TaskUs, an outsourcing firm that provides services to the crypto exchange. Another common tactic is social engineering attacks against outsourced IT and support desks, where threat actors impersonate employees and call BPO help lines to obtain access to internal corporate systems.

In one of the most prominent cases, attackers posed as an employee and convinced a Cognizant help desk support agent to grant them access to a Clorox employee account, allowing them to breach the company’s network. The incident later became the focus of a \$380 million lawsuit by Clorox against Cognizant.

Google reported that threat actors targeted U.S. insurance firms in social engineering attacks on outsourced help desks to gain access to internal systems.

Retailers also confirmed that social engineering attacks against support personnel enabled ransomware and data theft attacks. Marks & Spencer confirmed attackers used social engineering to breach its networks, while Co-op disclosed data theft following a ransomware attack that similarly abused support staff access. In response to the attacks on Marks & Spencer and Co-op retail companies, the U.K. government issued guidance on social engineering attacks against help desks and BPOs. In some cases, hackers target the BPO employee accounts themselves to gain access to the customer data they manage.

In October, Discord disclosed a data breach that allegedly exposed data from 5.5 million unique users after its Zendesk support system instance was compromised. While the company did not confirm how its instance was breached, the threat actors told BleepingComputer that they used a compromised account belonging to a support agent employed by an outsourced business process outsourcing (BPO) provider. Using this account, they downloaded Discord's customer data.

This repeated abuse of outsourced support providers shows how threat actors are increasingly bypassing vulnerability exploits and instead targeting third-party companies with access to corporate networks and data.

This is a variation on “the call is coming from inside the house”. In this case the call is coming from inside the house of someone you need to trust. The source of the inherent vulnerability is clear. In order for an external outsourced business process provider to perform their functions, they must be trusted with a connection into the outsourcing entity’s network or other business processes. Although they must be trusted, they are not worthy of that trust.

As I noted, an employee of an enterprise has an inherent stake in the company that employs them. They attend meetings with their fellow employees and look them in the eyes. They may socialize after work hours, attending each other’s birthday parties or those of their children. They may be on a softball team or have attended explicit team building events. They may share a department where they routinely meet, plan, participate and work side by side to meet goals. All of those things serve to create a stake in the shared welfare of the organization.

But none of that exists in the hearts and minds of subcontractors to whom that organization is just another account among many. This makes these subcontractors far more susceptible to bribery.

This newfangled restructuring of organizations appears to be irreversible. The days of an employee starting off in the mailroom and gradually working their way up over the course of five decades to finally receive a gold watch and become CEO are long gone and they’re not coming back. So how do we make this business process outsourcing work better?

My hope is that everyone is learning from these initial BPO missteps, and that the problems we’ve been seeing are due to API over-trust. In the same way it’s easier just to give someone wider permissions to a database than they need, it’s simpler and quicker to design an API that offers more power than is needed to fulfill an outsourced task.

For example, an external BPO which is providing helpdesk services may not need access to a customer’s entire record. They might only actually need minimum identifying information and a subset of specific customer history. But when initially setting things up, it’s quicker and easier to just give this “trusted” third party unfettered and unfiltered access to the entire customer database – after all, they’re under contract, right? What could possibly go wrong?

What we see is another example of the sort of finger-pointing I’ve been highlighting recently. Whose fault is it if a subcontractor is bribed to disclose their contractor’s critical information? The subcontractor is easiest to blame, but the information was still disclosed. The question is whether they had access to, and were this **able** to disclose, more information than was the bare minimum they needed to have to do their job? We need to start holding the designers of the interfacing APIs accountable for taking the easy way out and not taking the time to design and implement the disclosure of an absolute minimum of information to external service providers.

This “excess privileges” issue is not a new problem. BPOs were once called MSPs – Managed Service Providers. Years ago we covered the story of a dental services MSP which had been compromised by a ransomware group. This group struck gold because the way the MSP operated was to require full access to their client’s networks. The ransomware group took advantage of this unfettered network access to install ransomware and encrypt the PCs and other equipment of every one of the MSP’s customers. It was a widespread disaster for the MSP and for every one of the dental offices it served.

There was no defensible reason for the MSP to have a fully privileged network connection to each of its client’s internal networks, but that was the easy path that was taken. If the access had been strictly transactional against a service running on the client’s side, far less, if any, damage could have been done. So, philosophically, this is what must change. Any organization wishing to outsource services must consider the consequences of that service provider becoming a hostile entity. They must design and provide an API linkage that will protect their interests under any circumstances no matter what their contractors might do.

A familiar example of this sort of function is the HSM – the hardware security module – whose internal write-only private key and machinery can be employed to sign a file while at the same time nothing and no one can exfiltrate and steal its secrets. The analogy is not perfect, but the point I want to make is that designing with the concept of **least privilege** is what should always be done. Always. In the HSM example, there was no **need** to allow the device’s internal private key to ever be exposed – no matter how much the user of that key might be implicitly trusted. Thus the key should never be exposed, and never be exposable. Not because it would be stolen, but because it could be.

I’ve talked a lot about not exposing any non-public service to the public Internet. This is another example where “Least Privilege” comes into play. When I’ve said that authentication doesn’t work I’ve meant that it must not be **depended** upon it to work. I’ve asked why someone in North Korea, whom you almost certainly don’t intend to have accessing your enterprise’s network, should even be given the opportunity to challenge your network’s authentication system. If you were monitoring every incoming connection to the publicly exposed management interface of your enterprise’s firewall, and a connection attempt was unbound from North Korea, would you not choose to drop its packets? If North Korea is being allowed to connect to your cloud services, that’s not least privilege.

So my point is, even though the concept of least privilege could hardly be simpler and more easily explained – it is, afterall, a trivial concept – it is not trivial to actually deploy in every instance, so it’s still not something that is robustly deployed in the real world. But it needs to be.

I believe it’s the only way forward.

Through the years of this podcast I’ve broadly divided problems into two categories: (1) Mistakes that are made and (2) policies that are deliberate. AI-driven code-checking reasonably promises to finally enable us to deliver bug-free code. While that’s terrifically exciting, it won’t cure all of our ills because failures to implement “least privilege” are not mistakes, they are policies. They are the result of decisions. This means that to further improve our delivered security moving forward we need to make the decision to far more robustly design for least privilege operations.

