

Security Now! #1081 - 06-02-26

AI Captured the Flag

This week on Security Now!

- As expected, UnFiOS devices are under attack.
- CISA commands federal agencies to update Drupal.
- Can the largest botnet ever, be killed?
- Defender endpoint can cutoff a PC from the network.
- Charter Communications big account leak.
- Chrome moves device-bound session cookies from beta.
- Anthropic to release Mythos shortly.
- cURL and Daniel Stenberg.
- IBM and RedHat commit to fixing open source with AI.
- LOTS of terrific listener feedback this week.
- AI spells the end of a terrific source of training.

Not all signage seems necessary



Security News

UniFi Under Attack

I wanted to quickly follow-up on last week's urgent Ubiquiti announcement about their five identified vulnerabilities, three of which were maximum severity. The news is what we expected due to the widespread Internet exposure of more than 100,000 UniFi OS based devices – 50,000 of which are located in the United States: Attacks commenced immediately.

Last Tuesday afternoon, a posting on Linus Tech Tips noted that multiple users on Reddit were reporting that Unifi devices not patched for Ubiquiti's Security Advisory had a Super Admin user named "John Sim" added overnight, with additional users chiming in as their regions wake up. The attackers appear to exfiltrate data via the UniFi backup feature once inside.

u/k987654321: Hey guys can someone help me please. I'm away on holiday (in another country) and just had a notification that a Super Admin had been added to my account whilst I've been here. I logged onto the UniFi iOS app and there was someone called John Sim in there. I promptly removed it, as you can see.

u/thetoxicnerve: I just had exactly the same happen on my UDR. Same username too "John Sim."

u/jeffporten: Confirming that we saw the same attack, same username. We've removed the bastard from the superuser list and inspected the logs

u/EagerCDNBeaver: I also just had the same thing on mine.

u/thomasrw1: Just had 2 sites with this user created (have a lot more sites that were fine).

u/ravicc: I got hit with this too. I was on Unifi OS 5.0.16. I got the update notification on Thursday. I delayed the update last night since I was travelling this week. So, I was on the previous version of the UniFiOS. I also noticed that there were multiple backups triggered. Not sure where these backups went and what they were attempting to do. And what sensitive information is in the backups.

What then ensued over on the Linus Tech Tips forum was the typical back and forth about, among other things, whether automatic updates were a good thing. Everyone knows my feeling about that. Propeller heads want to be in the loop. We want to manage our own devices and decide for ourselves whether and when we wish to update. While that may have been practical ten years ago, it no longer is – at least not until our new AI code fixers have had the chance to give the entire industry's codebase a thorough going over.

Those arguing against enabling auto updates argue that a bad update might brick the router. They used the word "brick" to increase the drama of their position. But that use is inaccurate because "bricking" a device specifically means killing it beyond repair. All routers I'm aware of have the ability to revert to a known-working factory firmware image specifically to enable recovery from a bad firmware update – whether automatic or manual.

So no router is going to be "bricked" by any auto update failure. The worst that's going to happen is that a router won't boot after a bad update and will require manual recovery.

I'm not saying that's a good thing. But as I said last week, I now believe, given all of the evidence, that a mature manager, having weighed the risks versus rewards, will opt for enabling auto-update of their systems. Doing so will give them an extremely high probability of protecting their users from the attacks that are happening with increasing speed and frequency while having an extremely low probability of causing a network outage due to a failed update.

CISA: Federal agencies MUST immediately patch Drupal

Also last week we covered the critical PostgreSQL SQL injection vulnerability that affected pretty much all Drupal instances, even those that were very old and had not been receiving updates for years. This vulnerability was so bad that the Drupal team produced patches even for those long past end-of-life versions while leaving their many other security problems unpatched.

Against that backdrop we now have the U.S. government's CISA giving agencies one day to update. No excuses. Period. BleepingComputer has some nice coverage of this, writing:

CISA has given U.S. government agencies until Wednesday evening to secure their servers against an SQL injection vulnerability in the Drupal content management system (CMS) that it flagged as actively exploited. Drupal is typically used by large organizations managing massive data structures and multi-site installations, including government entities, educational organizations, major research universities, and high-profile enterprise and media organizations. Google/Mandiant researcher Michael Maturi discovered this vulnerability (now tracked as CVE-2026-9082) in Drupal's database abstraction API.

The security flaw can be exploited without authentication, allowing attackers to trigger arbitrary SQL injection on PostgreSQL-powered sites via specially crafted requests. Successful exploitation can potentially lead to information disclosure, privilege escalation, and even remote code execution. The Drupal security team tagged the flaw as "highly critical" before releasing patches and confirming that exploitation attempts had been detected in the wild.

Cybersecurity firm Imperva warned on May 21: "Since CVE-2026-9082 was released, Imperva has observed over 15,000 attack attempts targeting almost 6,000 individual sites across 65 countries. Attacks are primarily targeting Gaming and Financial Services sites so far, accounting for nearly 50% of all attacks."

Internet security watchdog group Shadowserver now tracks nearly 670 unpatched Drupal installations exposed online, most of them from North America (272) and Europe (273).

Friday, the U.S. Cybersecurity and Infrastructure Security Agency (CISA) added the flaw to its Known Exploited Vulnerabilities (KEV) Catalog and ordered Federal Civilian Executive Branch (FCEB) agencies to patch their systems by midnight on Wednesday, May 27, as mandated by Binding Operational Directive (BOD) 22-01. Although BOD 22-01 applies only to U.S. federal agencies, CISA advised all defenders, including those in the private sector, to apply CVE-2026-9082 patches as soon as possible to secure their organizations' devices.

CISA warned: "This type of vulnerability is a frequent attack vector for malicious cyber actors and poses significant risks to the federal enterprise [...] Although BOD 22-01 only applies to FCEB agencies, CISA strongly urges all organizations to reduce their exposure to cyberattacks by prioritizing timely remediation of KEV Catalog vulnerabilities as part of their vulnerability management practice. Apply mitigations per vendor instructions, follow applicable BOD 22-01

guidance for cloud services, or discontinue use of the product if mitigations are unavailable.” Over the last several years, CISA has flagged 5 Drupal vulnerabilities that have been exploited in the wild, two of which have also been abused in ransomware attacks.

We really do appear to be seeing a new world where no known vulnerability goes un-exploited.

Thank goodness there’s hope on the horizon with AI vulnerability discovery which can be employed **before** any new code is released. If this is done right, the phrase that has always rubbed me the wrong way – which is “all software has bugs” – can finally be provably refuted.

An astonishingly large botnet

During the third calendar quarter last year, Cloudflare was hit by and mitigated the largest DDoS attack ever reported. It clocked in at a wire-melting 29.7 trillion bits per second. This astonishing attack was attributed to the Aisuru botnet which Cloudflare estimated was composed of somewhere between 1 and 4 million infected hosting machines globally. Think of the scale of the task of assembling and managing somewhere between 1 to 4 **million** individual hosts which have all, one way or another, been collected and commandeered to serve a single master.

Because a botnet of 1 to 4 million already seems huge, I was surprised to learn that a far larger botnet was recently discovered by a security researcher who then reported the finding to the NCSC. After some additional investigation Dutch authorities in concert with the National Cyber Security Center in the Netherlands took down a set of **200 servers** that were being used to manage more than an astonishing **17 million bots** residing in infected hosts around the world. That’s right. A single managed botnet more than 17 million units strong. The reporting on this stated that: *“The police seized several botnet servers from a hosting provider for investigation purposes. The hosting provider then took the entire botnet offline because it was being used for criminal activities.”* Yeah, no kidding.

However, these days, botnets are not only being used to blast targets off the Internet by flooding and saturating their Internet connections. Today, there’s a large criminal demand for proxies. And that’s what appears to be going on here. The Netherland Times reported:

The Cybercrime Team of the Police Unit The Hague, together with the National Cyber Security Centre (NCSC), says it has successfully dismantled a large Asocks botnet and taken it offline.

The botnet was made up of at least 17 million compromised consumer devices around the world, including computers, routers, tablets, smartphones, and internet-connected devices such as smart security cameras. Investigators identified 200 servers used to run the infrastructure, all of which were physically based in the Netherlands.

The Asocks network operated as a “residential proxy service,” in which cybercriminals covertly infected poorly protected consumer devices with malware. These compromised devices were then used to route internet traffic and launch large-scale cyberattacks, all without the knowledge of their rightful owners.

The case was triggered by a report from a security researcher to the NCSC, which quickly passed the information on to the police. This led to a joint investigation by both agencies. During the operation, the Police Unit The Hague confiscated several servers from a Dutch hosting provider for forensic examination, while the provider itself shut down the malicious infrastructure once its criminal use was confirmed.

As consumer devices and routers are frequently targeted by proxy botnets, the police and the NCSC advise users to change default passwords right away, ensure their Wi-Fi is secured with WPA2 or WPA3, and install software updates as soon as they become available.

What's not clear, however, is whether the perpetrators of this comprehensive network failed to plan for this eventuality. My guess would be that they DID plan for this. If they were running an installation of 200 command & control servers for the purpose of managing the more than 17 million devices they had previously and painstakingly arranged to infect and control, then this was not some fly-by-night operation run by some random hackers. This would have been a serious commercial criminal enterprise. And my guess would be that it still is. I strongly doubt that simply shutting down the C&C servers will have been anything more than an inconvenience to these people and a momentary cash-flow interruption.

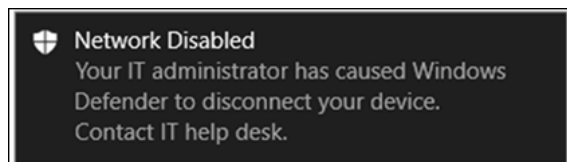
We've previously talked a lot about all the many ways disconnected bots can be rejoined to their command infrastructures. A favorite is to simply use an algorithmically generated DNS query. The individual 17 million+ members of the fleet that have been unable to reach their controllers at the previous DNS address, knowing the current time and date, will use an algorithm to synthesize a new DNS address which they will query to obtain the IP address for the updated command & control server infrastructure once it has been brought back online.

I'd say it's a pretty safe bet that this massive network of residential proxy hosts will be back in action just as soon as another willing hosting organization can be found and servers established.

No Internet for you!

Microsoft's May 2026 new features summary contains an interesting item. It's called "*Automatic device isolation (automatic attack disruption)*". It's currently in "Preview" status, but the brief description says: "*Microsoft Defender for Endpoint can now automatically isolate compromised devices as part of automatic attack disruption. Isolation blocks most network traffic while keeping the device connected to security services. The action is time-limited, scoped to the incident, and security operators can release isolation at any time.*"

Elaborating further elsewhere, Microsoft wrote: "*When a device in your organization is suspected of being compromised, Microsoft Defender for Endpoint can automatically isolate the device as part of automatic attack disruption. Automatic isolation helps reduce the risk of further impact on the organization, limit attacker lateral movement, and prevent impacts such as data exfiltration and ransomware propagation. When a device is isolated automatically: The compromised device is disconnected from the network, reducing the risk of further impact on the organization. But the device retains connectivity to the Microsoft Defender for Endpoint service, which continues to monitor the device.*"



There's much more information about this new feature that's now available in preview. I have a link in the show notes for anyone who is already deeply committed to Microsoft's solutions and whose enterprise might benefit from this automated compromised workstation isolation: <https://learn.microsoft.com/en-us/defender-endpoint/respond-machine-alerts#isolate-device---automatic-attack-disruption-preview>

Charter Communications Leaked 4.9 million accounts

Last week I received an email from Troy Hunt's "Have I Been Pwned?" notification service. It stated that GRC was affected by a recent breach at Charter Communications. While I value Troy's service, every time I've reacted to similar news and have taken the trouble to see which of GRC's accounts might be affected by a breach that HIBP captured and analyzed, I discover that it's a handful of email accounts that have never been valid. The grc.com domain has been around long enough and has acquired enough of the positive reputation that, unfortunately, its domain is used as a source of spam. And this is true even though anyone who might receive email claiming to be from GRC can trivially verify whether or not it was signed by GRC's server. Unfortunately, not all recipients bother to check even though SPF DMARC and DKIM have been well established for many years.

In any event, after receiving Troy's advisory, the news that there was indeed a sizeable breach at Charter Communications was not surprising, and neither was the news that the ShinyHunters gang was behind the breach. BleepingComputer wrote:

The ShinyHunters extortion gang stole personal information from 4.9 million accounts after hacking the U.S. telecom giant Charter Communications in early April, according to the data breach notification service Have I Been Pwned. Charter has over 92,000 employees and provides internet, mobile, video, and voice services to more than 32 million customers and over 57 million homes in 41 states across the U.S. through its Spectrum brand.

The company confirmed the breach earlier this week, saying that the attackers did not steal sensitive personal customer information and that it had alerted authorities about the incident. Charter told BleepingComputer: "No sensitive personal information (PI) or customer proprietary network information (CPNI) data was exfiltrated by the threat actor as a result of recent activity." While Charter has yet to attribute the attack and has not shared further details, the ShinyHunters extortion gang claimed responsibility and told BleepingComputer that they breached the company's systems on April 1 in a voice phishing (vishing) attack that compromised an employee's Microsoft Entra account.

The threat actors claimed they used this access to steal 42 million records from the company's Salesforce instance, including consumer and business customer names, email addresses, physical addresses, phone numbers, phone types, plan information, support ticket data, and some CPNI data. However, Charter spokesperson denied the gang's claims of CPNI data theft and said that "only sales tools used to manage current, past and prospective Business customers were impacted; no CPNI or sensitive PI was released by the threat actor."

After the company refused to pay the ransom demanded by ShinyHunters to have the stolen data returned and destroyed, the cybercrime group leaked the documents stolen from Charter's Salesforce instance on their dark web leak site.

Have I Been Pwned analyzed the leaked data and confirmed that the incident affected 4.9 million accounts, whose names, email addresses, job titles, phone numbers, and physical addresses were stolen. Have I Been Pwned said: "The group later published the data, which exposed 4.9M unique email addresses along with names, phone numbers and physical addresses. A subset of approximately 85k records originating from an internal employee directory also included job titles."

ShinyHunters has been targeting Salesforce customers over the past year, breaching hundreds of companies worldwide and claiming the theft of billions of records in Salesforce Aura data theft attacks and a Salesloft Drift campaign. The FBI has recently advised ShinyHunters' victims not to give in to the gang's ransom demands, after previously warning that doing so cannot guarantee that threat actors won't attempt to sell the stolen data to other cybercriminals or extort them again.

Charter Communications' systems were also compromised in a wave of breaches by a Chinese state-backed threat group tracked as Salt Typhoon that also impacted AT&T, Verizon, Consolidated Communications, Windstream, and Lumen, as well as telecom companies in dozens of other countries.

The unfortunate success of Voice Phishing, or vishing, attacks is a perfect example of the sort of cybercrime that AI used for software vulnerability discovery and remediation will not address. So even after our software is working the way we intend, we'll still have problems with security.

Chrome gains device-bound session cookies (DBSC)

We first talked about device-bound session cookies in detail nearly a year ago in July of 2025 when Google first announced their intention to support this technology. The name pretty much says everything. We saw way back in the days of the "Firesheep" Firefox browser extension that when HTTPS was only being used transiently during the privacy-sensitive logon – which, for example, is that Facebook was doing at the time – other people's session cookies after logging on could be easily captured and replayed to impersonate them in real time. This was possible because the cookie was a simple secret token that was assumed to remain the sole secret of the logged-on web browser. It was in no way "bound" – as in tied to or connected to – the physical web browser that had received that cookie from the web server.

So as I said, naming something a "device-bound session cookie" pretty much tells us everything we need to know. Google's Chrome browser has been testing this next-generation cookie tech in beta mode for some time, but last week they announced that it had moved into general availability. The technology allows browser cookies to be cryptographically locked to a single physical platform's TPM or Secure Enclave so that no one who might arrange to intercept it can successfully use it to impersonate its original owner. There is no user-facing behavior change. And as I noted at the time last summer, this does require extensive replumbing support at the server side. So it's unclear when that might happen for non-cloud-based providers. Someone like Google will deploy it and support it because they can. Our banks and social media providers may do so in some far off future. I'm not holding my breath for that one.

Anthropic to “soon” turn Mythos loose

Way down at the end of Anthropic’s May 28th announcement last Thursday, which announced their Opus 4.8 which replaces their previous Opus 4.7, under the innocuous heading “What’s Next?” they wrote:

Users will find Opus 4.8 to be a modest but tangible improvement on its predecessor. There’s still more to be done: we’re working on developing and releasing models that provide many of the same capabilities as Opus at a lower cost.

*Not only that, but we plan to release a new class of model with even higher intelligence than Opus. As part of Project Glasswing, a small number of organizations are currently using Claude Mythos Preview for cybersecurity work. Models of this capability level require stronger cyber safeguards before they can be generally released. We’re making swift progress on developing these safeguards and expect to be able to bring **Mythos-class models to all** our customers in the coming weeks.*

So, we have no date certain, and Anthropic is not elaborating on what the use of these new “Mythos-class models” will cost relative to Opus. But given all of the apparently well-deserved attention that Mythos has generated, it’s probably difficult to overestimate the demand Mythos’ release will likely create among companies whose current software offerings may be vulnerable to attack and who have not yet enjoyed access to Mythos, Daybreak or Codename MDASH.

cURL & Daniel Stenberg

We have an update from cURL’s Daniel Stenberg who, as we’ll recall, went through a grumpy phase induced by the overwhelming amount of “Ai slop” vulnerability reports he had been receiving. Then, more recently, ever since Anthropic’s announcement of Mythos in April, the entire software security industry has been seeing evidence that Mythos, while it was certainly great marketing on Anthropic’s part, was also much more than that. Last week we examined the impact Mythos had on Mozilla’s Firefox after they, in their own words, recovered from the vertigo of being hit with 271 vulnerabilities in code they had believed to have none. So what’s new with cURL’s author and maintainer, Daniel Stenberg?

Last week Daniel posted the following to LinkedIn:

Not even half-way through this #curl release cycle we are already at 11 confirmed vulnerabilities - and there are three left in the queue to assess and new reports keep arriving at a pace of more than one/day. 11 CVEs announced in a single release is our record from 2016 after the first-ever security audit (by Cure 53). This is the most intense period in #curl that I can remember ever being through.

So he’s already at 11 confirmed CVEs, while being not even half-way through the next planned release cycle with 3 reports still in the queue to assess and new candidates queuing at a rate of more than one per day.

Following this posting, Daniel added something to his own thread that I also wanted to share. He wrote:

The simple reason is: the (AI powered) tools are this good now. And people use these tools against curl source code. They find lots of new problems no one detected before. And none of these new ones used Mythos. Focusing on Mythos is a distraction - there are plenty of good models, and people who can figure out how to get those models and tools to find things.

Two things worth noting here. First, we know Mythos is good. It's the real deal. But it's also clear that Mythos is by no means the only game in town. Based upon what Microsoft has shared about "Codename MDASH", it sounds as if it might be another significant leap ahead of everyone else.

The second observation is what most interests me. Daniel wrote: *"the (AI powered) tools are this good now. And people use these tools against curl source code. They find lots of new problems no one detected before."* The first thing is that, just as Mozilla reported, AI is now discovering true vulnerabilities that have previously eluded humans. We know that while cURL has had its share of troubles, it has also been deeply scrutinized for many years – just like Firefox. But the second and most important thing to appreciate – which is also what Mozilla said last week – is that the problems are not infinite. There is some finite count of them and they are working toward bringing that to zero. Daniel is now diligently doing the same thing. Those 11 CVEs he already has are resolved. They are fixed. So he will be approaching a time when no one is able to find anything else wrong. And cURL, like the rest of the industry's software which has gone through the AI wringer, will be demonstrably more correct and secure.

Project Lightwell:

IBM along with RedHat just announced their "Project Lightwell" with joint commitment to spend \$5 billion to help find and fix vulnerabilities in open-source software packages. They plan to deploy more than 20,000 engineers with AI tools as part of this new project. Their initial focus will be the Maven and Java ecosystem. It will then expand to PyPI, npm, Go, and others.

So it appears that the open source world will have some angels to help foot the AI-bill to clean up its latent vulnerabilities. This is great news.

Listener Feedback

AI

I got a kick out of this piece of listener feedback. A listener whose name I recognize from his years of occasional feedback wrote:

Hi Steve, I love your podcast, but it is getting to be too much AI. Speaking of AI, do you think there might be some way to enlist AI to stop robo calls?

What made me shake my head a bit was AI's expressed annoyance over this podcast spending so much time on AI, but then in the same breath, asking my opinion about an AI-related matter. As I noted last week, I'm certain that this wall-to-wall podcast-consuming coverage of everything AI is transient. But when I step back to examine how much we've all learned through this podcast's coverage about what's going on right now, which is nothing short of a massive transformation in the way complex software is authored and made far more correct, I cannot imagine having spent any less time looking at these changes.

I recall once feeling similarly when we were spending a lot of time examining the nature of the first early ransomware attacks back when they were a novelty. My feeling then was that they represented a significant pivot in the world of cybercrime, and as we now know years later, most of the cybercrime ever since has been about exfiltration and extortion.

So my point is, I believe that our listeners are being well served, even when I may appear to be spending undue time on something. My spidy-sense is telling me that we are again in the midst of another significant pivot, one of the biggest ever.

Travis Hayes — re: Velocity of Vulnerabilities

Hi Steve, I'm enjoying catching up on this week's SN and your thoughts on how AI is starting to gain real traction in finding (and patching!) vulnerabilities. It seems like we are seeing the beginning of a huge increase in the supply of quality vulnerability hunting, which leads me to remember all of those "supply and demand" curves they tried to teach me about in Econ 101 when I was in college. Since the supply is ramping up, I am curious on your thoughts regarding where we will reach equilibrium on the demand side. While there are multiple drivers, I'm thinking specifically about two.

First: bug bounty programs. Large companies and organizations have been funding substantial bounties, motivating clever people to work hard to earn big, juicy rewards. These programs have been huge advertising wins for many companies-- a large cash commitment has the effect of convincing consumers that the company is serious about security, adding confidence to buyers and gaining headlines when high dollar amounts are rewarded.

Second: Zero-day hunting, either in contests like Pwn2Own, or directly to black-market buyers.

With the huge increase of relatively cheap ability that AI agents are poised to bring to the table, it seems to me that the motivation (i.e. demand) for these activities is going to dry up quickly. Why would Microsoft or Google continue to offer 5-figure bounties for threat hunting,

when they will be able to do at least as good, or better, of a job themselves in house? The black market for exploitable zero-days should also collapse, no?

Thanks for your continued insights and instruction; I always look forward to my weekly visits with Uncle Leo and yourself. -Cheers, /Travis

I was also thinking about the effect of all this on things like Pwn2Own. The only thing that makes sense to me is that – at least initially – there may be corners of the software industry that do not get around to employing AI-based software quality assurance. So they would represent ripe targets. But it's not clear whether any human researcher would be able to outperform emerging machine intelligence. If it turns out that AI is equal to humans – and let's be 100% clear about the fact that everyone who works with Mythos, for example, comes away with that conclusion – then I agree with Travis that Bug Bounty programs and Pwn2Own are very likely to go the way of the dinosaur. They will become memories of the way things once were done, like using punched cards and paper tape.

Joseph Fienberg — AI certification?

Steve, Three decade listener to SN. I was an AI skeptic, but you changed my mind. You gave me hope at the end of SN 1080 that making code bug free is possible. But, I believe third party certification that code is AI tested for vulnerabilities any time a change is made may soon be required by market forces. No one will sell me a toaster that is not UL listed and safe to plug in. When I use software, or visit my bank's website, I want certification that their entire ecosystem of code has been independently AI tested. I will gladly upgrade to Windows 12 if Microsoft certifies that a neutral third party AI tested their code.

What's going to force this is that businesses will be contractually required by their insurance companies, banks, customers and suppliers to "AI certify" their products are bug-free. A real world example would be businesses accepting credit cards would have to certify as a condition of accepting credit cards that they AI tested their systems. Regards, Joseph, St. Louis, MO

That's a truly interesting spin and it's not something that had occurred to me. I can see the logic behind it. Until now, software was a somewhat mysterious art. It was a "best effort" where all anyone could do was hope for the best. But now we have systems that are able to autonomously demystify the code humans have created and are able to give it a gold star, a blue ribbon, or a formal certification.

A perfect example is the before and after effects of Mozilla's Firefox which we covered last week. Today, its code could be certified as having passed "The Claude Mythos vulnerability analysis." Once upon a time, Mythos found 271 vulnerabilities. Today, it finds none – not one. So that can actually mean something significant. It's a concrete assertion.

We know that insurance companies are inherently risk averse. They already require things like annual security audits and assertions that all of one of their insured client's edge systems are running the latest firmware and are current on patches. If a breach then occurs which a company attempts to file an insurance claim for, if those earlier representations can be shown to have been fraudulent, that can be grounds for denying any claim.

Joseph's observation is that until now, due to the inherently unfathomable nature of software, there was no means for making any sort of meaningful assertion about the provable quality of software. But now Mythos, Codename MDASH, and presumably someday Daybreak are demonstrating that they have the ability to "fathom" the arbitrarily complex systems we humans concoct. That being the case, I'd say that Joseph's notion of software certification by AI is quite likely to occur.

Adam Merkley — Claude to the rescue

Hi Steve, I work for an MSP here in the Phoenix area, and I wanted to share a quick win I had recently using Claude. One of our customers needed to swap out an aging Fortinet FGT-60E for a new Unifi gateway. I suspected the network was fairly flat, but anyone who's spent time in Fortinet's UI knows how easy it is to lose the will-to-live, scrolling through those menus — so I wasn't exactly looking forward to auditing the config manually.

Instead, I exported a configuration backup from the Fortinet appliance and fed it into Claude. I asked it to summarize every configured setting and map each one to its Unifi equivalent. Within seconds, I had a clean, actionable breakdown: a defined WAN IP, a LAN subnet scope, and a handful of port forwards — nothing more. Claude not only confirmed my suspicions about the unknown Fortinet device's setup being basic, but told me exactly what to configure in Unifi for a clean drop-in replacement.

What would have been a tedious, error-prone manual review turned into a two-minute task. I was genuinely impressed. /Adam Merkley, Scottsdale, AZ

I loved this note from Adam. It's a perfect example of the power of this new Genie. Those of us who are actively using it are discovering new uses for its capabilities every day. But at this point adoption varies widely. Here's what's going on: Throughout our lives we have all built up a model of the way things work in the world. We know how things work. And for the most part, nothing much changes from day to day. But then, almost overnight, everything actually did just change.

Some people have not yet awoken to that fact, and "a fact" is what it is. I get it that not everyone has experienced this dramatic change in the world. And if the nature of your life and work is not helped by having mostly accurate nearly instantaneous linguistic access to most of the world's knowledge, then perhaps AI won't ever impinge upon your life. And that's fine too.

But everyone who has been following this podcast knows that I am anything but an early adopter. I mean, come on, I'm still programming in assembly language! And I will soon be reluctantly giving up Windows 7 and updating GRC's use of Server 2008R2. I held onto my TiVos until they stopped working and I was forced to give them up. So when I, Steve Gibson, reluctant adopter of newfangled things, excitedly disclose that I have discovered, and now have an active working partnership with, an AI named Claude which is allowing me to be vastly more productive in my daily work, I hope our listeners will appreciate that the world really has significantly changed.

About now, Adam, who wrote that note, is nodding his head knowingly. He realizes that AI provides him with a form of leverage he's never had before. And, like me and so many others of

us, he's still discovering the endless new things that we've always accomplished by ourselves on our own, and tend to continue to, out of sheer lifelong habit. Those old habits are now outdated because we are suddenly able to express these needs and questions to an over-eager machine assistant who stands by, almost too helpfully willing, to assist us. Everything has changed. I am 71 years old. I am not collaborative by nature. By my own choice I have always worked alone. But I now have an enjoyable working partner. It may be a bit weird, but it's real.

Steve Meyers — Here's another example

His email Subject was: "*Claude helped me switch from EdgeRouter/X to a Ubiquiti UCG Ultra*"

I bought the UCG Ultra last summer, but never actually did the upgrade, because my network is complex enough that I expected it to be painful. Last month, I decided to see if Claude could help me out with the upgrade. I've found that it's very good at making plans -- for my software projects, I'll usually have it create a plan, then we'll go back and forth on some of the details before I have it assist me in implementing the plan.

So I asked it to make me a plan. I gave it a config backup from my EdgeRouter/X to work with, and it created a Python script that would transfer the DHCP, port forwarding, and firewall rules to the UCG. I also transferred my backup from my Unifi Controller for my access points to the UCG as part of the process. It gave me a step-by-step plan, including what to do offline, how to do it, and how long each step should take.

The firewall rules transfer did not work, because the new UCG routers changed to a zone-based firewall, so I had to do that manually. That wasn't a huge deal, but it showed how it isn't perfect. The work was all done with the UCG offline so it didn't affect my network at all. I've enjoyed hearing about your exploits with Claude, and thought I'd add a little anecdote about ways you can use it besides just coding. /Steve.

I'm going to stop myself from sharing more of these sorts of emails because this is what our listeners are discovering — they are coming to realize that they no longer need to do everything themselves. In Steve Meyer's case, he had already been using Claude to assist with planning software projects, but for the past year — since last summer — he had been putting off the pain of switching network routers due to the need to translate his existing network configuration to a new and different router. Then it occurred to him that he might be able to ask Claude to help. And help it did. He gave Claude a config dump from his current router and it wrote a Python script to configure his very different replacement router similarly. Thus, most of the pain was sidestepped.

Frank S. — Is all this knowledge collection dangerous?

And this leads us nicely into another listener's story and very important question:

Steve, I too have built what seems like a relationship with Co-pilot (my AI selection) and it knows an enormous number of things about me now. I've been using it hard since August of 2024. At that point in time, it was helping me finish my Equilibrium Pro App. (Currently In Windows Store) I've watched the behavior change as the client side went from browser use to the actual Co-pilot App. I told my Bot that I wanted it to set itself to remember the maximum possible, and since having said that, it has collected an enormous amount of information about me. It knows about my code, and my home network in very precise ways. It knows about my

interests, and even my style of conversation. It's astounding to see what it's capable of doing, and how personalized it feels for me. At some point I suspect the free ride will be over, and Microsoft will come calling for payment. It will be very hard to let it go. It's become a part of my work flow now, in almost anything I'm doing.

So, this brings me to a question I'd like to hear about on the show: What happens when a bad actor finds a way to impersonate me, and talk to Co-pilot as if they were me? I shudder to think how much information they could glean from such a move. They would know a great many things about me, my home network, my code base etc. They would also know a great many things I'm interested in. How would I even know the bad actors were doing it?

Kindest regards - Frank S.

And in the same vein...

Joshua Krichman

Hi Steve, I'm curious about your thoughts on the struggle I'm having with diving head first into AI. I'm a systems architect and engineer in a small org trying to wrap my head around the idea of allowing an AI hosted in the cloud by any of the major vendors, to know almost everything about me. Currently I tell my users that even though my org pays for AI tools where the data isn't used for training on the vendors models and that the data is housed in our own pod, be careful of what kind of information you feed into it. The more data housed in the vendor, the more opportunity there is if there's a breach of some kind in the vendors infrastructure.

If that's the advice I'm currently giving my users, how can I not take my own medicine? I'd like to personally start using AI as I can clearly see the benefits it could have not only on my work life, but my personal life as well. But I'm reminded that the only way AI will work, is if I feed it more and more specific and/or personal data to have it tailored to my liking. Taking into account that most companies aren't worried about users privacy and security and much more about their own bottom line, what's the best approach here so I don't get left behind? This is one of the main reasons why I host most of my personal data with Apple rather than Google. Any thoughts you have on this would help guide me in the right direction.

I've been watching Leo since the days of TechTV and have been a listener of SecurityNow for many years. Thank you for continuing past 1000!

Frank and Joshua bring up very good points. There's a very clear potential downside to all of this cloud-centric long-term user context accumulation – which has grown to be a major factor in the use, value and success of today's AI.

My one-year subscription to [Venice.AI](#) expired last Wednesday, so I know it was one year ago that I discovered it and shared that discovery with our listeners. At the time, I played with it a bit to see whether it was truly uncensored. And I can affirm that, yes indeed-dee, it will happily converse about and produce images of anything one might ask. But after taking it out for a test spin, I tired of it and decided I had no particular need for an uncensored AI. To Venice's credit, they gave me ample notice and warning of my annual subscription renewal. That brought me back there for the first time in a long time to look around. One of the things that caught my attention again was Venice's affirmation that they store **ALL** of their users' gradually

accumulating context in the user's local browser and NONE of it on their remote servers. I mention this in the context of Frank and Joshua's notes, which I'm sure echo what many of our listeners are feeling and may be concerned about. I can well imagine that some of our listeners might be more than somewhat put off by the idea that the AI they are becoming quite chummy with and with which they may be choosing to confide increasingly deep and personal aspects of themselves, might someday be breached. Since there's an aspect of "the more you give the more you get", users who choose to contribute more of themselves are rewarded, much as Frank noted, with a significantly more personalized experience. It doesn't take long before one's resistance to such sharing is overcome.

Add to this the fact that many of the major players have less-than-perfect security records themselves. So far, four supply-chain incidents have hit OpenAI, Anthropic, and Meta. None of these targeted the AI models themselves; all four exposed the same gap: release pipelines, dependency hooks, CI runners, and packaging gates. But it doesn't inspire confidence in the security of a (currently) cloud-based service that very much wants to know as much as about as it can. We worry about data brokers compiling our various stats and credit bureaus leaking our social security number, physical address and birthday, and while our AI assistants might not know any of that, they tend to wind up acquiring and deliberately retaining and digesting a huge amount of deeply personal information.

If these things ever do evolve into advertising-supported services – as perhaps the free services will – they will have more absolutely accurate advertising targeting information at their disposal than any advertiser could ever dream of. It would put Google to shame.

Another aspect of this that's worth mentioning is contextual knowledge lock-in. Every AI service has its own internal bespoke representation for the knowledge it has accumulated about its users – and this knowledge is non-portable to other services. It's true that it's possible to perform a poor-man's transfer by asking an AI to please display everything it knows and then feeding that into another service. But the loss of information fidelity makes this barely worth the trouble. And it's unclear why any of these services would be interested in developing such a facility, if it were even possible. Since they have each invented their own schemas for ingesting and digesting, it's unlikely that we're ever going to see that. So, as any one service's knowledge of us grows over time, the tendency to remain "loyal and faithful" to that AI will also grow. Dare I say that it's analogous to chatting with an old friend who already knows you well versus striking up a conversation with a stranger at some cocktail party who asks: "So what do you do?"

Though I never used [Venice.AI](#) enough to know whether it offers personalized context equal to what ChatGPT, Claude or the others do, it might be worth exploring that if you have become concerned about how much personal information has been accumulated by your current AI service.

Our long-time listener, **Sabrina Tarson**, has another perspective:

Hi Steve, I've been listening to this podcast on and off since Episode #256 (LastPass) back in 2010 when I was still in high school. I was very lucky to start having the time to listen to the podcast again when the news dropped about Project Glasswing and Mythos.

I have to say, it's very refreshing to hear both you and Leo's experiences with AI. Most of my generation (I'm in my early 30s), and the generation younger than myself, are completely anti-AI to the point where they swear to never touch it with a 10 foot pole. Like you, I feel this is a very shortsighted view, and one mainly born from ignorance. Hearing how two people use it in their daily lives, for helping get work done, who were around when all this tech around us was just starting get off the ground (the personal computer, cellular phones, the internet, etc) is not a perspective I get to listen to often, and one that I deeply appreciate.

My only concern with these AI models, a larger concern that I have about my career's future (I'm a sysadmin at a small company), is mainly the companies that are currently running them. I trust in the technology, it is the future, no matter what people's opinions on it are. It's the ultimate Pandora's box, and it's never going away. At the same time however, the major frontier models are created and operated by some of the largest corporations imaginable, and unfortunately, their end goal is monetary. We've seen it time and time in the tech industry, first it's innovation, then it's about the quickest way to harvest our data and sell it to the highest bidder. OR, in the case of shady organizations like Palantir, use the data our AI's learn about us ultimately for control, working with a corrupt government as we currently live in today.

My hope, is that eventually this dependence on massive corporations to run these models is reduced, and the AI we're going to need for cybersecurity and our own personal lives, are localized, on-device models that are either powerful enough to run on our phones or computers, or for us nerds, compact enough to run a server at our own homes, keeping our own data private. These AI are learning about us every day, helping improve our workflows, but ultimately, they are owned by these massive corporations, and the industries track record for handling our personal data gives me pause.

I'd be interested to hear both yourself and Leo's thoughts on this. Thank you both for a wonderful podcast :)

Infatuated and astonished as I am by how much faster I am able to move forward with Claude quickly extracting for me extremely specific and detailed knowledge from the global knowledge pool, I nevertheless want more control.

More than anything else I want to have this running in something that might resemble a quietly humming NAS box in a closet. This thing would have whatever local processing and storage it needed, along with a connection to the global Internet. Just like automobiles, these would be available in a range of "models" with the higher-end choices delivering their answers faster and probably also incorporating more knowledge. The concept of "model size" expressed in trillions of parameters – perhaps "Terameters" – would become commonplace. What was once "how fast is your Internet?", "how large is your screen?" or "how fancy is your car?" would become "how many Terameters is your home's AI?" Those who don't mind waiting longer for an answer, or who may not also wish to use the services of autonomous agents, could get by with the economy AI package. And, of course, simply using an online cloud account will always be the low-investment option.

But with such a device quietly humming in the closet, the various members of my household – in my instance my wife Lorrie and me – would be separately known to it and readily identified to it

by the various devices we use throughout the day.

Another intriguing possibility is that a hybrid local-cloud relationship might evolve. Imagine that our local AI box retains all of the user-specific knowledge of us. That's where all of the personalization occurs and where all of our various "agents" live. In this model, a great deal can be done locally. But there might be a need for our local AI to occasionally reach out when it needs to have some sort of heavier lifting done on it or its users' behalf. In that case, the local AI would protect the privacy of its owners by making generic requests for information from the big daddy cloud AI.

Another strong case for having a locally operating AI is that it seems clear that the next huge win for AI will be the creation of autonomous agents that are continuously working on our behalf behind the scenes in the background. We don't appear to currently know how to do that safely, but we're going to figure that out because it's clearly too powerful for us not to. I would tell my AI to be sure to let me know when Peter F. Hamilton, Ryk Brown or some other of my favorite authors release any new Sci-Fi, and also when any streaming Sci-Fi that it thinks I might like becomes available, but not to notify me until all of a season's episodes have been released because I prefer to binge. Please pay my monthly bills, let me know specifically if anything varies by more than 10%, email me a monthly summary of costs and accounts, and so on.

All of this is clearly coming. Given that local models are already showing viability for various tasks, and that we have barely begun to explore and understand these new capabilities, I have absolutely zero doubt that there will be an Apple HomeWise AI device and devices made by companies that have traditionally manufactured home NASes, routers and similar appliances. It's going to happen.

AI is going to follow the same trajectory we've seen with all previous technologies, but I suspect that the pace, which is already breathtaking, is not going to slow down. Although I doubt I'll purchase Apple's HomeWise AI box, since I'll prefer to build my own, I fully expect to be doing so within the next ten years.

So, yes, Sabrina, with you currently in your early 30's, I really would wager that by the time you're in your early 40's – and likely well before then – not only will AI be deeply entrenched into our lives, but we will also have many cost-effective local solutions to choose among.

Brian Weeden — Open Source access to non-free AI tools

Listening to your comments about Mozilla's response to Mythos, I think you have it right that this will ultimately be a good thing for those developers who have the resources and time to fix their code. But that makes me wonder if this might be disastrous for the developers who don't have resources or time, namely the open source community.

Mozilla has a dedicated team of security professionals to wade through the Mythos results and fix things, and they have time to do so because their code is closed source. But I don't think major open source projects have either of those.

First, one correction: Mozilla's Firefox is 100% open source. They were one of the organizations

within Anthropic's Project Glasswing who received early access to Mythos specifically because of the strong perceived need for a publicly exposed project like Firefox to be made as secure as possible. In the short term, Anthropic said that as part of Project Glasswing it will be providing \$100 million in usage credits and \$4 million in direct funding to support open-source security efforts. So that will be some help. But there's a much bigger point that I want to make...

Reusing the historical mass storage analogy because everyone can relate to it and because it's also so apropos... back when the first IBM PC's were appearing, a 10 megabyte hard drive cost five thousand dollars. A few years later you could get 40 megabytes for the same price. Today, one million times more storage, which would be 40 terabytes, costs far less and is incalculably faster.

My point is that nothing about the economics of today's AI will be true tomorrow. If I'm sure of anything it's that AI is going to follow a similar technological development curve and collapse in cost. So while today, yes, leveraging the capability of AI for the creation of bug-free code is not free, yet it is already entirely feasible and cost effective for commercial software publishers. There's no question in my mind that ten years from now it will not only be widely available but also completely taken for granted.

Lisa Lombardo — re: "The Burroughs"

Hi Steve, Thank you for the recommendation of "The Burroughs." It's kinda like Stranger Things for Boomers. I like the engineering of Sam and his daughter. PS. Plus the Boss and other great music. /Lisa.

Lisa and others enjoyed the pointer, so I'm glad that I thought to mention it. And Project: Hail Mary continues to be finding many fans around the world and among our audience.

AI Captured the Flag

I am extremely sensitive to the fact that so much of this podcast has recently been focused upon AI and its impact upon our lives, our privacy and security. But this is the **Security Now!** Podcast and the impact that large language model artificial intelligence is having right now across the entire spectrum of the computer security industry could hardly be more relevant.

So I want to conclude this week's podcast by sharing the text of a terrific blog post written by a security researcher named Kabir Acharya. Kabir introduces himself on his "About" by writing:

Hi! I'm Kabir, a senior security engineer with a deep passion for highly technical pentesting and security research. I spent my time at Atlassian applying Application Security concepts to modern technologies including LLMs/AI, networks, AWS/GCP/Azure cloud platforms, SaaS integrations and in-house products and tooling. Now I work at Transgrid, securing Australia's largest electricity network and its OT environment. I play CTFs on the global stage with Emu Exploit, HashMob, and TheHackersCrew and produce music in my spare time.

Doing a bit more digging, we learn that during Kabir's last six months of his four years with Atlassian he conducted more than 250 security reviews and supported software and Machine Learning engineers to make better security decisions. He delivered more than 15 security threat models, improved understanding of information risk in platforms including Forge and Rovo (which are AI/LLM-based), he found, reported and aided patching of more than 10 security vulnerabilities external to threat models, and patched more than 70 security vulnerabilities. So this guy clearly lives security and he's no stranger to its growing intersection with large language models.

In his auto-bio he wrote: *"I play CTFs on the global stage with Emu Exploit, HashMob, and TheHackersCrew"*. While we have previously spent a great deal of time looking at the Pwn2Own competitions through through the years, somehow we haven't before focused upon CTF's, which stands for "Capture the Flag" competitions. These are very popular hacking contests where participants solve security-themed puzzles and challenges to find hidden strings of text — which are the so-called "flags" — that are then submitted to a scoring system for points. These competitions are one of the primary ways people in the security community learn, practice, and demonstrate offensive and defensive skills within a legal, structured environment.

These CTFs range from beginner-friendly events hosted by university clubs to elite international competitions. DEFCON's CTF, which is held annually during the DEFCON conference in Las Vegas, is considered to be the pinnacle, often referred to as the "World Series" of CTF due to the pedigree of the competition's participants. Teams must qualify through preliminary events to compete. Other well-regarded competitions include Plaid CTF (run by Carnegie Mellon's PPP team), Google CTF, picoCTF (which is designed for high school and college students), and many dozens of others tracked on sites like [CTFtime.org](https://ctftime.org).

Whereas the Pwn2Own competitions are discovering original vulnerabilities, the CTFs are about discovering planted secrets. Both serve important roles for the industry. The CTFs are a legal sandbox for practicing techniques that would be illegal to use against real systems. And they

have served as a recruiting pipeline — top CTF performers are heavily recruited by security firms, intelligence agencies, and tech companies. They build the shared culture and vocabulary of the field. And they often produce write-ups afterward, where teams publish detailed explanations of how they solved each challenge, creating a corpus of freely available security education. Many of the researchers who discover major real-world vulnerabilities got their start (or stay sharp) through CTF competition.

Okay. So we have some sense for who Kabir is, and also for what CTF competitions have traditionally meant for the industry and to those who wish to use them to sharpen their hacking skills in a competitive environment. And we know that Kabir has been one such person having participated in a number of CTF competitions and teams.

Kabir's blog posting is titled: "*The CTF scene is dead.*" The details are very interesting and many are very insightful and important. He wrote:

Frontier AI has broken the open CTF format. The scoreboard does not measure human skill cleanly anymore, and the old game is not coming back. What makes me qualified to say this?

I started playing CTFs in 2021, the same year I started university. My first CTF was HCKSYD, a 48-hour solo CTF. I fully solved it and won in 2 hours. I was completely hooked. That led me to win DownUnderCTF, Australia's largest CTF, with team Blitzkrieg multiple times. Blitzkrieg was one of Australia's strongest teams at the time. I later joined TheHackersCrew, an international top-tier team that was consistently ranked highly on CTFTIME, the main global ranking and event calendar the scene uses as its scoreboard. With them, I competed in some of the most prestigious CTFs in the world, consistently placing well within the top 10 until the end of 2025.

I am not saying this because I dislike CTFs. I am saying it because CTFs were the thing that made me fall in love with security. They taught me how to learn, gave me a way to measure myself, and introduced me to many of the people I respect most in the field. Watching people pretend the format is still fine is frustrating because the old game is not there anymore.

What changed? *As AI tools ramped up in capability, especially when GPT-4 first came out, a significant percentage of medium-difficulty CTF challenges started becoming one-shottable, meaning a single prompt from a user could produce the solve and flag. You could paste a cryptography challenge into ChatGPT, come back in 10 minutes, and have the solution. At the time, we did not think too much of it. Hard challenges went mostly untouched, and the time save was not large enough to ruin the competition.*

The issue was never that AI could help. CTF players have always used tools. The issue is when the model does the reasoning, writes the solve, and leaves the human with nothing meaningful to do besides copy the flag.

Enter Claude Opus 4.5: *When Opus 4.5 dropped, the tone changed. Almost every medium difficulty challenge, and some hard challenges, became agent-solvable. Claude Code packaged everything into a CLI and made it easy to connect other CLI and MCP tools. It became trivial to build an orchestrator that used the CTFD API to spin up a Claude instance for every challenge. You could let the system run for the first hour, then only start working on whatever was left.*

That changed the game. Teams that refused to use AI were not just missing a convenience; they were playing a slower version of the competition. Open online CTFs started becoming a

question of how quickly you could automate the easy and medium work, then how much human attention you had left for the hardest challenges. The scoreboard started measuring orchestration and willingness to use frontier models along with, and sometimes instead of, security skill. The effects were obvious. The CTFTime leaderboard started feeling wrong. Some legendary teams that were consistently near the top appeared less often. Player activity felt lower. Challenge developers who treated CTFs as an artform had less reason to spend weeks building something beautiful if it was going to be eaten by an agent in minutes.

Then GPT-5.5 sealed the deal: I have been working heavily with GPT-5.5 and GPT-5.5 Pro after launch. By benchmark metrics, 5.5 is close to Claude Mythos' capability, and Pro likely surpasses it. These models can one-shot **"Insane"**-level difficulty active, leakless, heap pwn challenges on HackTheBox. They can solve a large portion of what a smaller CTF organiser can realistically produce. If you orchestrate Pro against Insane challenges in a 48-hour CTF, there is a good chance you get the flag before the event ends.

That makes open CTFs **pay-to-win**. The more tokens you can throw at a competition, the faster you can burn down the board. Specialised cybersecurity models like alias1 by Alias Robotics are becoming less relevant compared to general frontier LLMs. The competition is turning into "who can afford to run enough agents, with enough context, for long enough."

CTFs feel much more like a cheesable mess than a competition. Your performance in a CTF no longer defines your skill the way it used to. Recruiting security practitioners by CTF performance is becoming less meaningful. It is not even a particularly good measure of AI skill, because most of the orchestration needed for CTFs is already open source or vibe codeable.

The "beginners are fine" take: I have seen various takes that beginners can still learn from CTFs as they always have. These takes miss the scoreboard. CTFs were not just a set of puzzles. They were a ladder. Even as a beginner, you had something to climb. You could see yourself improve, solve more challenges, place higher, join better teams, and become more competitive over time. That feedback loop is breaking. If the visible scoreboard is dominated by teams using AI, a beginner is pushed toward using AI before they have built the instincts the AI is replacing. **That is an anti-pattern.** It prevents active learning, and active struggle is the bit that actually teaches you. It is also completely demotivating to put in real effort and see no visible progress because the ladder above you has been automated.

It also changes what challenge authors want to build. If beginner CTFs become another place where people quietly paste prompts and climb a scoreboard, authors have more reason to put their effort into learning platforms instead. At least on platforms like picoGym and HackTheBox, the expectation is education, and beginners are less incentivised to cheat themselves out of learning. Beginners are better off using picoGym, HackTheBox, and other lab environments where the point is actually learning instead of pretending the public scoreboard still reflects human growth.

So, "CTF isn't dead" ?? I have seen some hopium posts about how CTF is not dead, it is just augmented by AI. They often point at CTFs like DEF CON to argue that AI still cannot solve everything. That is true, but it is the wrong defence.

The hardest top-tier finals have very few participants, and they are usually gated behind qualifiers that are easier than the finals themselves. If those qualifiers fall to agents, fewer genuinely qualified people reach the challenges that still resist AI. A tiny number of elite finals does not save the open online format that most people actually play.

The claim is not that every challenge is solved. The claim is that enough of the scoreboard has been automated that the scoreboard no longer means what it used to mean.

What about the "AI is useful for security research" take?

CTFs were never meant to be security research. They can showcase new and interesting techniques, but the CTF itself is not the point of discovery. Just because AI is useful within a field does not mean it belongs in the competitive landscape of that field.

In CTFs, unrestricted AI removes the human from the puzzle almost entirely and reduces the art of security — to a prompt. Sure, LLMs will keep getting better at security as long as CTFs are around, but that does not mean the competitive format is healthy. CTFs were an artform, a way to share techniques with nerds, and a way to push the human bounds of security skill. That purpose is being stripped away.

What about the "LLMs are chess engines for cyber" take?

Chess has been dominated by computers for well over a decade. People use chess engines as an analogy for LLMs in CTFs, but they miss the point: chess engines are not allowed during competitive play. They are used for analysis, training, commentary, and practice. They enrich the game around the competition without replacing the person competing.

Imagine giving every competitive chess player the best chess engine and letting them use it freely during matches. Would that be considered fair? Would it be fun to watch? Would it justify prize pools? Would it push the human limits of what could be achieved in chess? The same questions apply to CTFs.

And organisers cannot fight back.

CTF organisers have tried techniques to break or deter LLM solutions, but they are temporary friction at best. Claude Code does not meaningfully care about old refusal-string tricks anymore. Frontier models are getting better at noticing prompt injections. Web search capabilities weaken challenges based on technologies released after the training cutoff. Rules that ask people not to use LLMs are ignored and almost impossible to enforce in open online events.

That leaves organisers in a bad position. If they make normal challenges, agents solve too much. If they make challenges deliberately hostile to agents, the challenges often become guessey, overengineered, or unpleasant for humans too. That is not a real fix. It just makes CTFs worse for everyone.

So "just adapt bro"

This take is infuriating. People I have always looked up to in the community have said it. To me, it is completely nonsensical unless you explain what we are adapting into. If adaptation means building better tooling, CTF players already did that. If adaptation means writing harder challenges, organisers already tried that. If adaptation means accepting that the scoreboard is now an AI orchestration benchmark, then we should say that honestly instead of pretending the old competition still exists.

Even if organisers create guessey or more overengineered challenges that current LLMs cannot solve, there are no good paths for players to learn the required skills while staying competitive. A few models from now, that point may be irrelevant anyway. The trajectory of LLM security capability is moving too quickly for challenge design to stay ahead for long.

So what's the aftermath?

The scene that grew my love for CTFs is emptying out. The CTFTIME leaderboard has almost no semblance of history or human skill anymore. The 2026 scoreboard is unrecognisable compared to every year before it. TheHackersCrew, alongside many other large and reputable teams, either do not play, play with far fewer people, or struggle to cut into the top 10. Unregulated cheating is through the roof. Some of the best CTFs, like Plaid CTF, are not running anymore.

*These sentiments are not only mine. Many members of my local team, Emu Exploit, feel similarly. These are people who consistently attend the International Cybersecurity Championship, perform at the top level in bug bounty programmes, compete in Pwn2Own, and present at conferences including Black Hat. The people losing interest are not casual observers. **They are exactly the kind of people the scene used to produce and retain.***

The fun of CTFing is gone for many of the people who cared most. The loss is not just a scoreboard. It is the ladder from beginner curiosity to elite competition. It is the craft of challenge design. It is the feeling that a clever human solved something difficult because they understood it deeply.

That legacy is not being carried forward by open online CTFs in their current form. The format is dead. Something else may replace it, but pretending nothing fundamental has changed only makes the loss harder to talk about honestly. It also gives AI skills more room to capitalize on the decline by selling mediocre wrappers back to the community that made the training data valuable in the first place.

So what now?

While a lot of what's happening in the CTF/AI space is super commercialised and out of our control, CTF has had a hugely positive impact on the industry. I have met so many kind, smart, and passionate people through CTFs. I have played some of the most beautifully crafted challenges and found some of the most intriguing unintended solutions.

The community around CTFing has been an amazing place to learn, grow, and connect. That's something we should not lose, no matter where the competition goes. As a community, we should strive to stay together and build new avenues to stay passionate and keep learning. Security-adjacent social events like SecTalks, student conferences, and local meetups are great ways to stay connected and stay involved. Learning platforms and the communities they provide through platforms like Discord are also a valuable resource.

While it may be a struggle to find an alternative to what we had, the amazing community we have built around it is more important now than ever as we find new ways to keep the competitive spirit alive.

I believe that we all need to deeply understand, appreciate and internalize that the entire field of software security – as we have always known it – has been forever changed this year. Mozilla knows this. Daniel Stenberg knows this. Kabir, who is mourning the death of his beloved and supremely valuable capture the flag competitions knows this. Pwn2Own will die. The software bounty industry will dry up. And, bless their hearts, “Zerodium” – the firm that purchases 0-days for resale to undisclosed dark parties – will also blessedly die.

All of the many various enterprises that have been built up over time, as a side effect response to the fact that we have been able to easily create software systems that were too complex for us to fully understand ... are likely to fall into the dustbin of history.

It is what it is. There's no point in mourning it. It's going to happen because now that, for the first time ever, we **can** have secure software, that's what we're going to have. What we are **not** going to have going forward is anything that exists solely because software has vulnerabilities.

