WELCOME TO THE 11TH ANNUAL HACKING CONFERENCE

Check-In Code: dojo

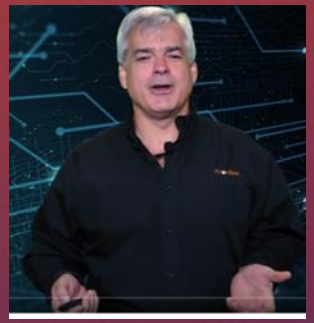
The Institute of Internal Auditors Chicago

ISACA Chicago Chapter

knowbe4

Rise Above Risk

How AI Impacts Patch Management



Roger A. Grimes, CISO Advisor
e: rogerg@knowbe4.com

Roger A. Grimes

CISO Advisor, KnowBe4

39 Years+ in Cybersecurity, 8+ years at KnowBe4

QUICK BIO

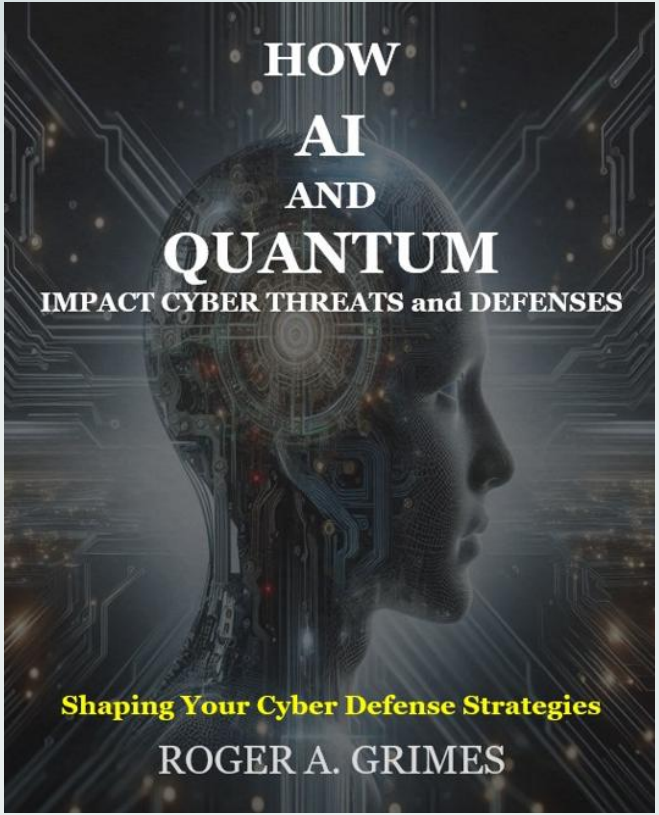
- Started fighting computer viruses in 1987 with John McAfee
- Previously worked for Microsoft, Foundstone, and McAfee
- Previously VP of IT, Network Manager, 20 years pen testing, cybersecurity instructor
- Consultant to world's largest to smallest companies plus militaries for decades
- Career goal - wants to "Fix the Internet"
- Written 16 books and over 1,600 magazine articles

SOCIAL PRESENCE

- Over 42K LinkedIn followers
- 5.4k X followers
- Writes 1-3 original articles a week
- Posts 1-3 times a day
- WSJ, CNN, Newsweek and NPR's All Things Considered
- Member of Cloud Security Alliance, WEF, MIT Forum

50+ CERTIFICATIONS

- CPA, CISSP, CISM, CISA
- MCSE: Security, MCP, MVP
- CEH, TISCA, Security+, CHFI
- BS in Business Administration (Accounting major/Econ minor)



https://www.amazon.com/dp/B0GGB7Y855/

Free PDF download: https://www.knowbe4.com/hubfs/AI-Quantum-Book-Roger-Grimes.pdf

About KnowBe4

We help ~70,000 organizations build a strong security culture to manage the ongoing problem of social engineering and human risk

BEST SOFTWARE 2025 Top 100 BEST SOFTWARE PRODUCTS

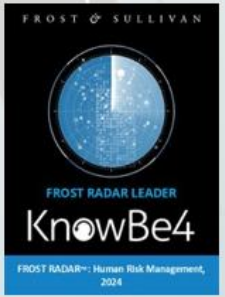
Global Sales, Courseware Development, Customer Success, and Technical Support teams worldwide

Gartner Magic Quadrant Leader

Trusted by 47 of the world's top 50 cybersecurity companies, and the largest human risk management platform

TrustRadius Top Rated 2025

CEO & employees are industry veterans in cybersecurity



Offices in the USA, UK, Canada, France, Netherlands, India, Germany, South Africa, United Arab Emirates, Singapore, Japan, Australia, and Brazil

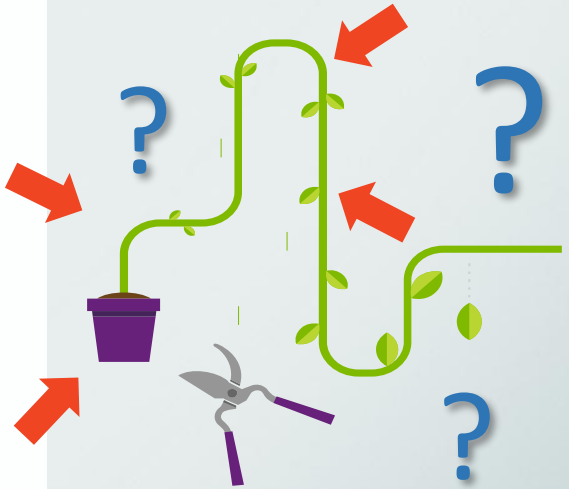
Agenda

- Current State
- What Really Is AI?
- Impact on Vulnerabilities
- How You Need to Update Your Patch Management Program

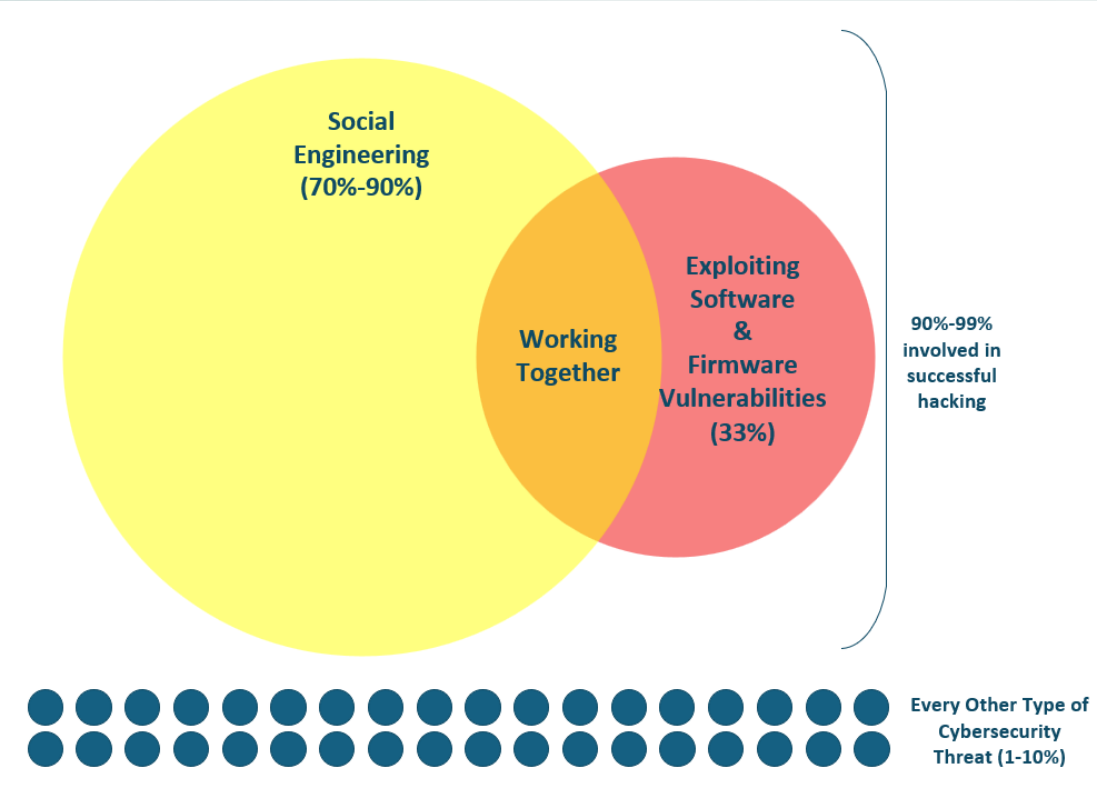
Current State

How Hackers Hack

- Social Engineering
- Software or Firmware Vulnerability
- (Technical) Impersonation/Authentication Attack
- Intentionally Malicious Program/Instructions/Scripting
- Human Error/Misconfiguration
- Eavesdropping/MitM
- Side Channel
- Information Leak
- Brute Force/Computational
- Data Malformation
- Network Traffic Malformation
- Insider Attack
- 3rd Party Reliance Issue (supply chain/vendor/partner/etc.)
- Physical Attack



Most Popular Initial Breach Root Causes



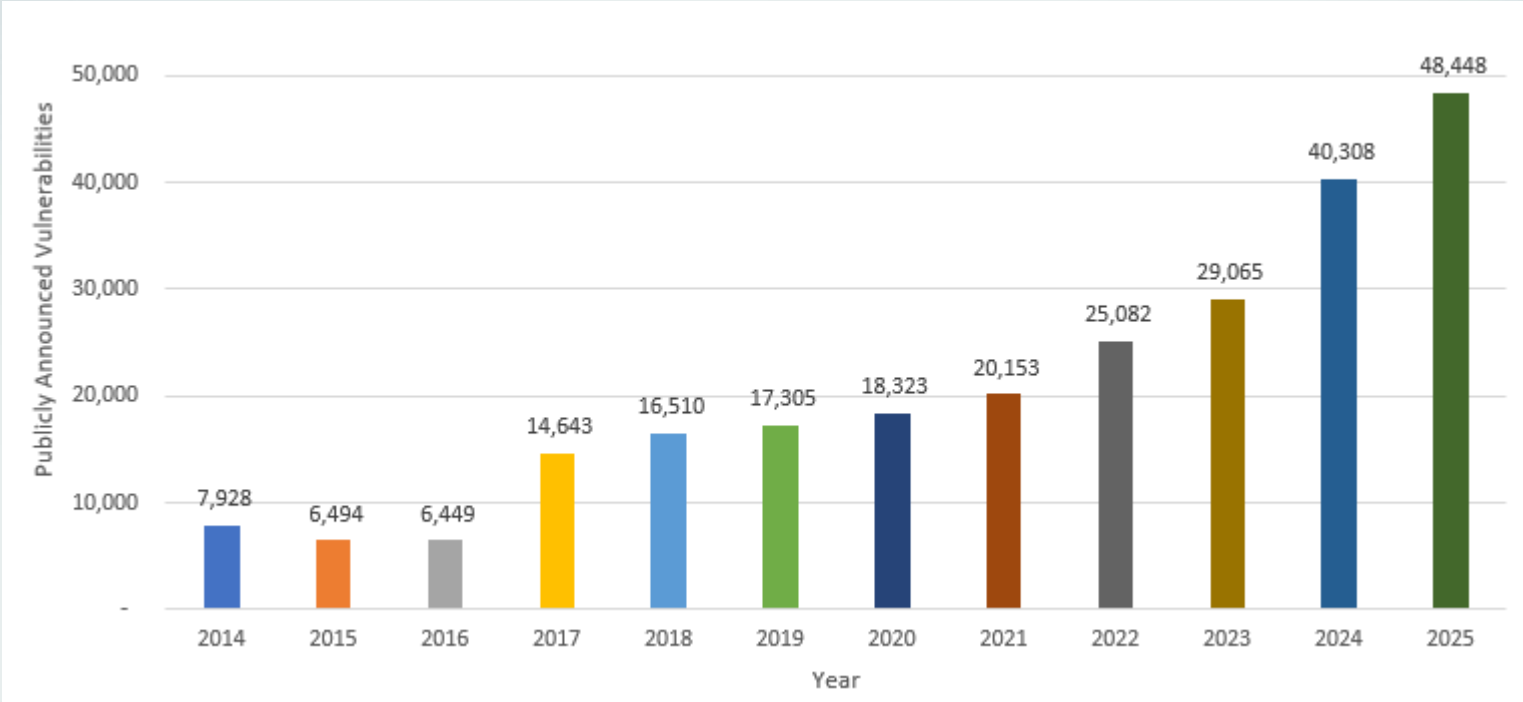
Software and Firmware Vulnerabilities Account for 33% - 40% of Successful Attacks

https://blog.knowbe4.com/hands-on-defense-unpatched-software-causes-33-of-successful-attacks

Tens of Thousands of Vulnerabilities Per Year

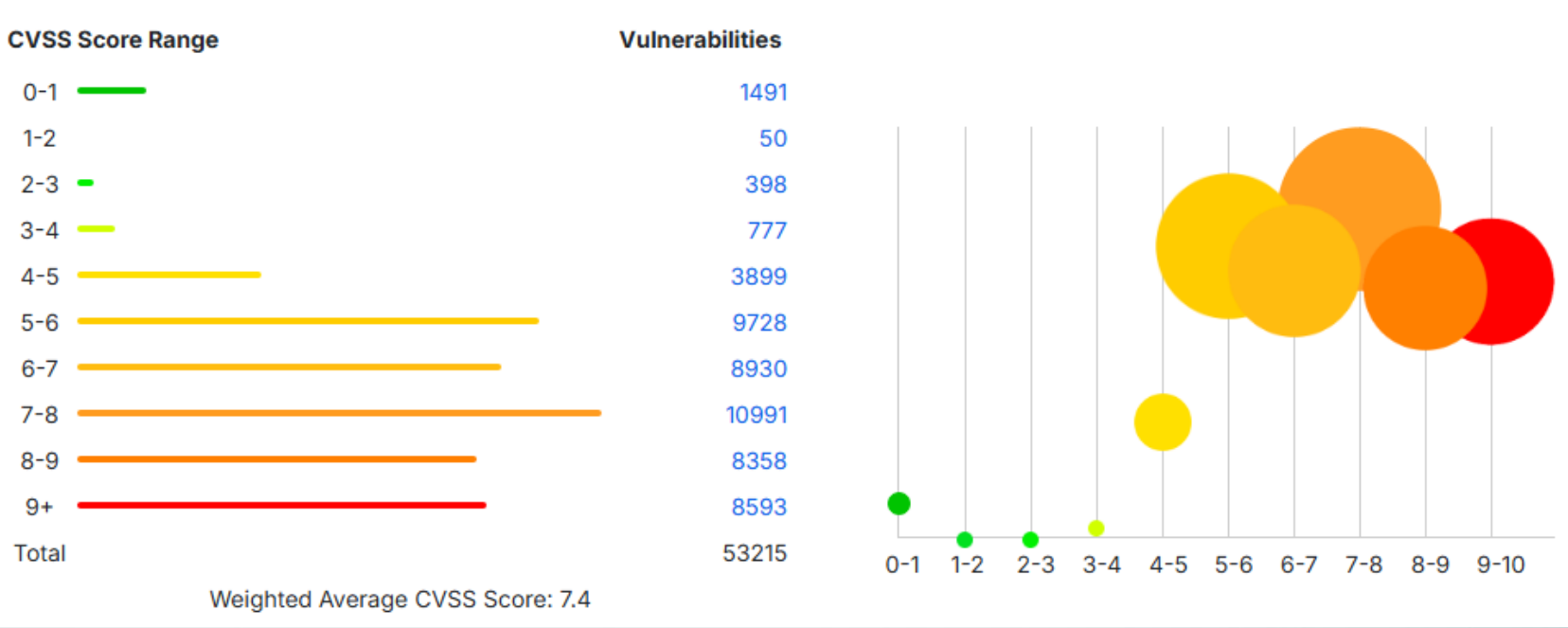
- 48K+ new threats in 2025
- More than a 132/day, accumulating day after day

And this is just publicly known vulnerabilities



Most are High Criticality

- Most vulns are ranked 7 or higher on a scale from 0-10



https://www.cvedetails.com/cvss-score-charts.php

Less Than 1% Exploited In Real World

- Luckily, most announced vulnerabilities are not exploited

Year	Total CVEs	Exploited	% exploited
2022	25,082	130	0.52%
2023	29,065	163	0.56%
2024	40,308	163	0.40%
2025	48,488	180	0.37%

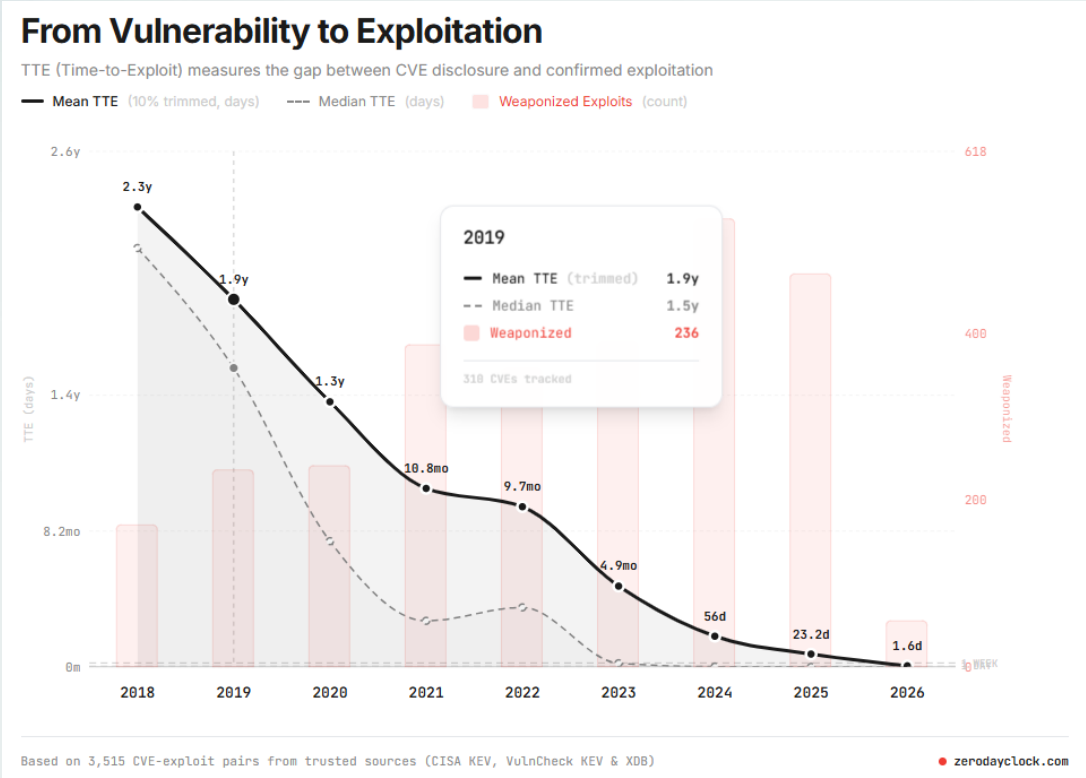
https://www.cisa.gov/known-exploited-vulnerabilities-catalog

Vulnerabilities Are Exploited in Days

- **If exploited, mean time from CVE to first exploitation in wild is 1.6 days**

https://zerodayclock.com/

**Trend is clear,
it will get worse**

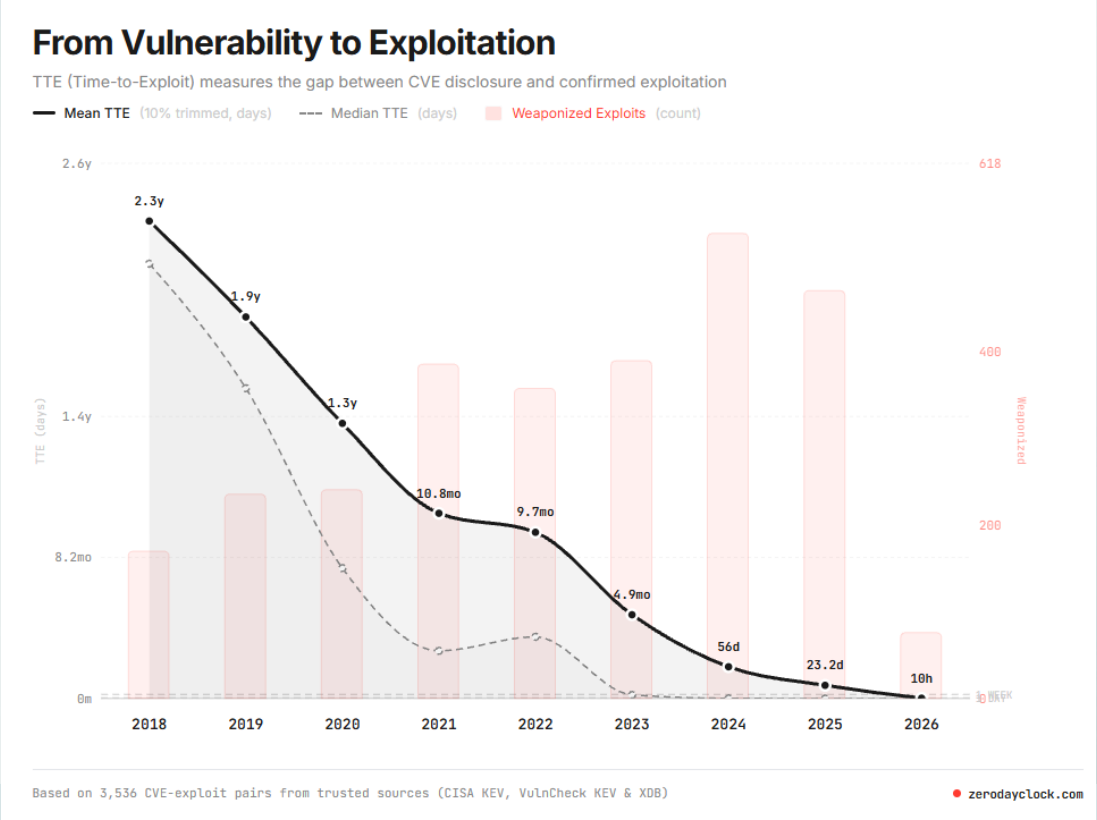


Vulnerabilities Are Exploited in Hours

- **Whoops…in just a few weeks it's down to 10 hours**

https://zerodayclock.com/

**Trend is clear,
it will get worse**



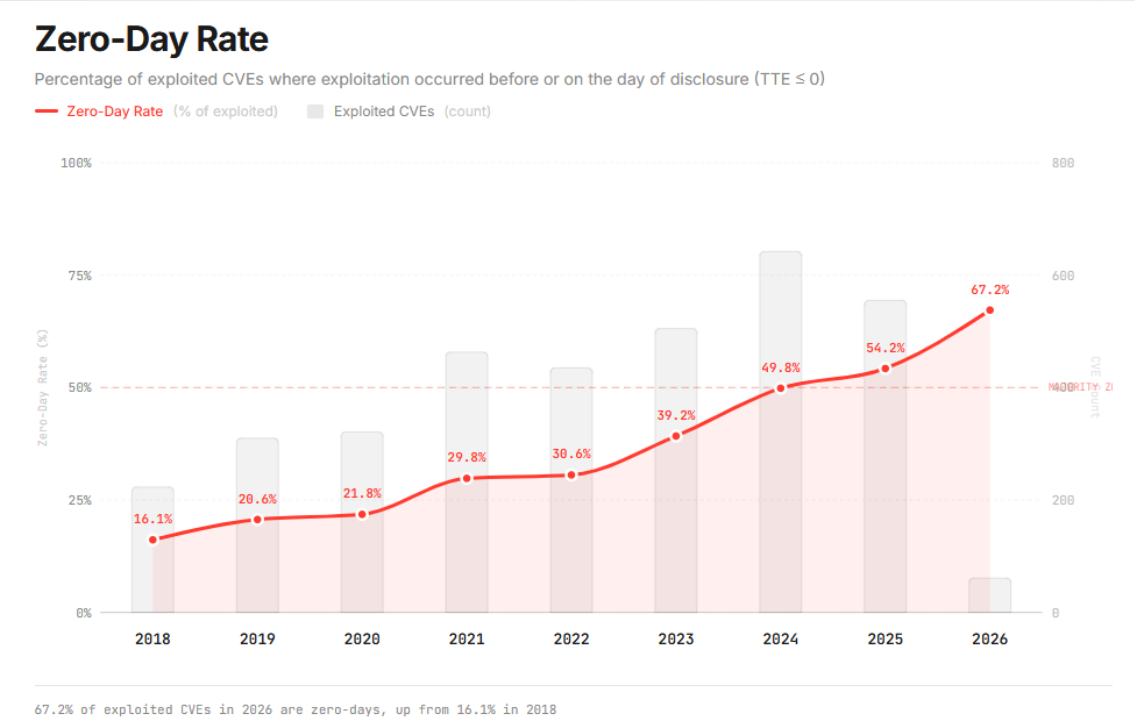
Most Vulnerabilities Are Zero Days

- **Zero-days (0-days)** are vulnerabilities not yet known to the public and/or vendor for which a patch may not yet exist at time of public announcement

Became that way in 2025

https://zerodayclock.com/

Trend is clear, it will get worse



Faster, But Not Fast Enough Patching

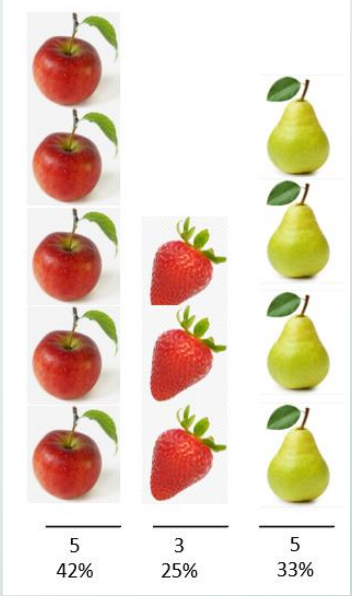
- Average time from patch availability to patching has been declining for decades
- Today, the (mean) average patching speed is 1 week after the vendor releases patches
- Popular software is patched faster than less popular software
- Software is patched faster and more often than firmware
- 20% - 24% - percentage of vulnerable instances are not patched in the first year or ever

What You Should Already Be Doing

What Really Is AI?

What Really Is AI?

- A general-purpose, probabilistic pattern-matching engine
- It takes data in, finds patterns (i.e., “tokens”), calculates the likelihood of pattern, and outputs results based on query (i.e., prompt)



What Is Agentic AI?

Agentic

General Definition

- **Software/service that uses separate, stand-alone, but cooperating “modules” to meet a common goal**

What Is Agentic AI?

Agentic

General Definition

- **Software/service that uses separate, stand-alone, but cooperating "modules" to meet a common goal**

- Real-world allegory: Like building a house

  (construction manager, plumbers, electricians, concrete people, roofers, painters, flooring, inspectors, etc.)

What Is Agentic AI?

Agentic AI

General Impact of AI

AI Attacks

Doesn't do everything magically better

But already excels at:

- Pattern-Matching
- Automation
- Speed
- Pervasiveness
- Predictive vs. Reactive
- Remembering (context)
- Hyper-Personalization

Impact on Vulnerabilities

AI Attacks

2 Major Categories of Attack

- Attacks from AI Against You
- Attacks to the AI You Use

AI Attacks

Attacks From AI

Traditional Attacks Made Better+

- Far Better Social Engineering
- Deepfakes
- OSInt Bots
- Hackbots
- Better Bruteforce Attacks Against Non-Random Targets
- Hallucinations
- Unexpected Emergent Behavior
- Data Leaks
- Data Theft
- Privacy Invasions
- Tool Mis-Uses

AI Attacks

Attacks Against AI

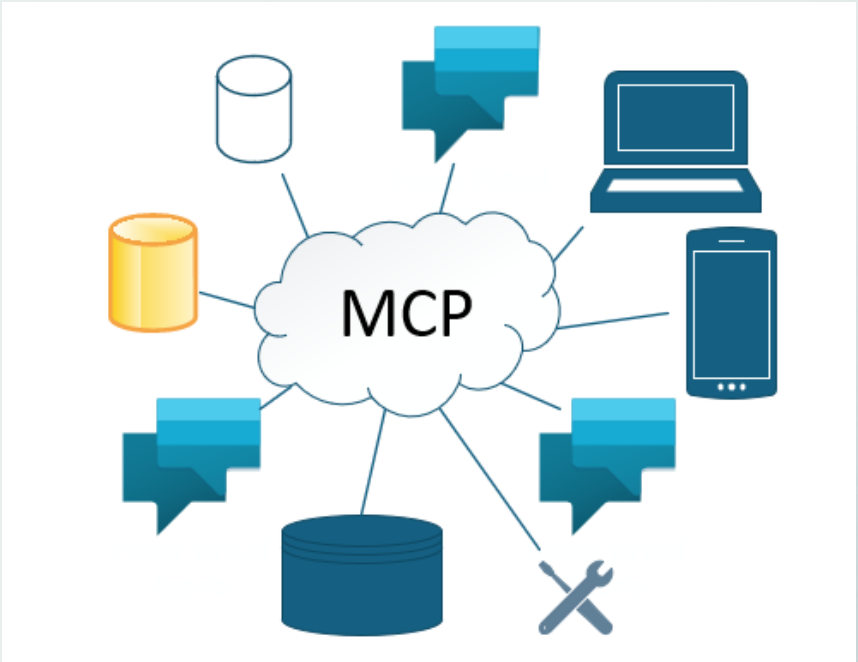
- Prompt injections
- Data poisoning
- Context poisoning
- AI identity attacks
- Supply chain attacks
- Jailbreaking
- Abusing AI system prompts
- Model/weight manipulation
- Label poisoning
- Memory poisoning
- Improper input handling
- Improper output handling
- Excessive agency
- Unbounded consumption
- Attacks against AI browsers
- Attacks against AI-browser add-ins
- Privacy risks
- Ad-driven attacks
- API attacks
- MCP attacks
- A2A attacks
- Malicious models
- and more

AI Attack Example

MCP Attacks

Model Context Protocol

- MCP connects AI, AI agents, tools, and data
- Is and will be hugely attacked



AI Attack Example

MCP Attack Example

1. You're running Claude Desktop AI agent on your desktop
2. It reads, summarizes, and helps you with email
3. Attacker creates a prompt injection attack and sends it in email to you
4. Your AI intercepts email, "consumes" it, and the attacker gets full access on your system

RealWorld MCP Attack Examples

Critical Vulnerability in Anthropic's MCP Exposes Developer Machines to Remote Exploits

CVE-2025-6514: A flaw in the popular mcp-remote proxy tool lets malicious MCP servers execute arbitrary commands on a client's machine.

CVE-2025-49596: A backdoor in Anthropic's official MCP Inspector tool allows RCE due to a lack of authentication, turning any developer machine into a target.

MCP stacks have a 92% exploit probability:

Filesystem Vulnerabilities: Two critical bugs (CVE-2025-53110 & CVE-2025-53109) in the widely-used Filesystem MCP Server permit directory traversal, letting attackers read or write files anywhere on the system, far outside intended sandboxes.

The Postmark MCP server attack: 5 key takeaways

A malicious Model Context Protocol package was found in the wild last week. Here are lessons from the compromise of the AI interface tool.

AI is Finding Lots of Vulnerabilities

Anthropic Mythos

Mythos Preview has already found thousands of high-severity vulnerabilities, including some in *every major operating system and web browser*. Given the rate of AI progress, it will not be long before such capabilities proliferate, potentially beyond actors who are committed to deploying them safely. The fallout—for

During our testing, we found that Mythos Preview is capable of identifying and then exploiting zero-day vulnerabilities in every major operating system and every major web browser when directed by a user to do so. The vulnerabilities it finds are often subtle or

Anthropic's New A.I. Model Sets Off Global Alarms

Mythos has triggered emergency responses from central banks and intelligence agencies globally, as Anthropic decides who has access to the powerful model.

Mythos: An AI tool too powerful for public release

by Pieter Arntz | April 20, 2026

AI is Finding Lots of Vulnerabilities

Anthropic Mythos

During
exploit
browse

then
major web
subtle or

too
c release

This is actually a good thing as you can use AI to find and fix your vulnerabilities faster

AI is Finding a Lot of Vulnerabilities

But More Bad News

- AI enables more attacks which require great complexity to accomplish
  - Attackers can more easily and frequently implement harder-to-exploit attacks
  - Attack complexity is no longer a barrier
  - Result: Attackers will be more successful, more often
- Patching and testing takes time, so it creates a permanent negative asymmetry for defenders

AI is Creating a Lot of Vulnerabilities

Unsafe At Any Speed

- AI is allowing more coding faster
- So far, AI (vibe coding) is making as many vulnerabilities as humans did
- Just based on sheer lines of code with consistent vulnerability rates, we will have more vulnerabilities
- Maybe, one day, AI will create fewer vulnerabilities in the code it writes, but we are not there yet

AI is Creating a Lot of Vulnerabilities

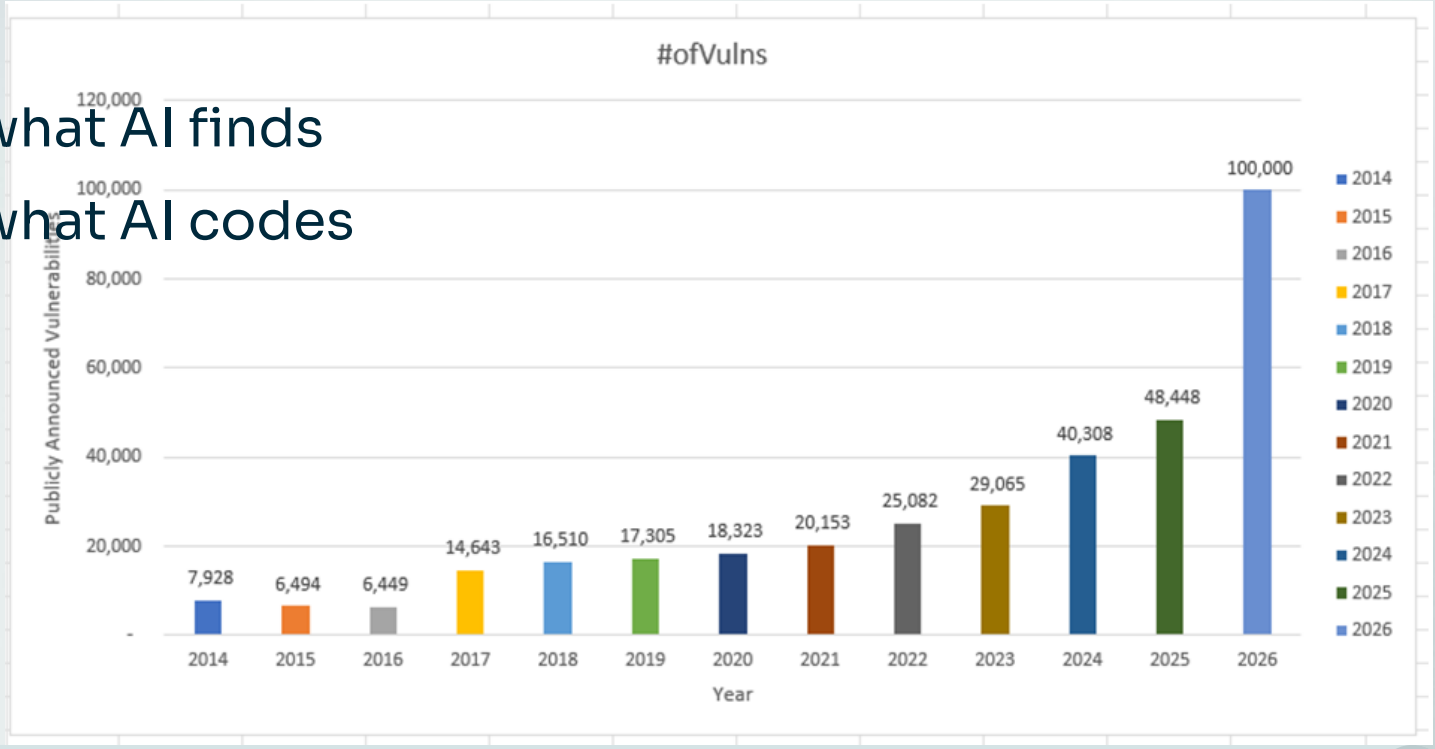
Prepare for a Massive Wave of Patches

- A ton of patches will be coming
- Patching workflows could become overwhelming
- Stress, anxiety, burnout coming your way

Vulnerabilities in the Age of AI

I Predict Over 100K Vulns This Year

- Half from what AI finds
- Half from what AI codes



Fully Agentic AI Attacks

A hacker used AI to automate an 'unprecedented' cybercrime spree, Anthropic says

The company behind the Claude chatbot said it caught a hacker using its chatbot to identify, and extort at least 17 companies.

But the case Anthropic found is the first publicly documented instance in which a hacker used a leading AI company's chatbot to automate almost an entire cybercrime spree.

According to the blog post, one of Anthropic's periodic reports on threats, the operation began with the hacker convincing Claude Code – Anthropic's chatbot that specializes in "vibe coding," or creating computer programming based on simple requests – to identify companies vulnerable to attack. Claude then created malicious software to actually steal sensitive information from the companies. Next, it organized the hacked files and analyzed them to both help determine what was sensitive and could be used to extort the victim companies.

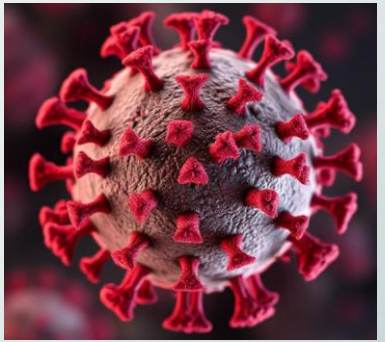
The chatbot then analyzed the companies' hacked financial documents to help determine a realistic amount of bitcoin to demand in exchange for the hacker's promise not to publish that material. It also wrote suggested extortion emails.

https://www.nbcnews.com/tech/security/hacker-used-ai-automate-unprecedented-cybercrime-spree-anthropic-says-rcna227309

Malicious Agentic AI Deepfake Threats

Overview

- One day instead of being human-led…
- They will just run themselves
- Morph and update code and capabilities as needed
- Contain its own vulnerability scanner
- Have its own break-in engine
- Be able to become who it needs to be
- Be able to move from A-Z efficiently



- This isn’t a possibility…this is what will happen

AI Hacking Bot

Everything hackers do today, but automatic, better, faster, changing as needed

Highly Autonomous Malware

Vulnerabilities in the Age of AI

Even Good Guys Practicing Responsible Disclosure and Pro-Active Patching Will Decrease Patching Cycles

"Responsible disclosure" cycles are decreasing

1. AI will see the patch
2. Figure out the vulnerability
3. Make the exploit
4. Start sploiting

Vulnerabilities in the Age of AI

Patching Will Decrease Patching Cycles

Example – Copy Fail

- Linux kernel vulnerability responsibly disclosed on April 29, 2026, 5 weeks after initial discovery
- Allows any unprivileged user to gain root access on every major Linux distribution using a 732-byte Python script with 100% reliability
- Vulnerability existed since 2017
- Took a good guy 1 hour to find with AI
- Hundreds of millions of Linux distributions not patched when disclosure was made

https://byteiota.com/copy-fail-cve-2026-31431-732-bytes-to-root-on-all-linux/

Vulnerabilities in the Age of AI

Patching Will Decrease Patching Cycles

Examp

- Linux k… 026, 5 weeks…
- Allows… or Linux distrib… ty
- Vulner…
- Took a…
- Hundr… disclo…

```python
# Simplified attack flow (pseudocode)
alg_socket = socket(AF_ALG, SOCK_SEQPACKET, 0)
alg_socket.bind(("aead", "authencesn(hmac(sha256),cbc(aes))"))

# Splice target file into crypto socket
splice(target_file_fd, crypto_socket_fd, offset)

# Trigger AEAD operation → writes to page cache
recv(alg_socket)  # Corrupts /usr/bin/su in memory

# Execute corrupted binary
execve("/usr/bin/su")  # Root shell
```

https://byteiota.com/copy-fail-cve-2026-31431-732-bytes-to-root-on-all-linux/

Vulnerabilities in the Age of AI

Patching Will Decrease Patching Cycles

Example – Copy Fail (discoverer released because of fear of AI discovering patching bits)

Brian Pak @brian_pak · May 1

We (Theori / Xint team) reported the vulnerability privately to the Linux kernel security team on March 23. We worked with them on patches, which landed in mainline on April 1. CVE assigned April 22. We disclosed publicly on April 29 with a full write-up and PoC.

Brian Pak @brian_pak · May 1

We spent significant resources (people, time, tokens) to help secure the kernel. We reported responsibly, helped with the fix, and waited for the patch to land before going public.

Brian Pak @brian_pak · May 1

Additionally, once the patch was already in mainline for a while and the CVE was assigned, we became concerned that AI systems monitoring commits could automatically detect the bug's criticality. So we moved to disclose promptly.

Brian Pak @brian_pak · May 1

This is exactly why we believe the entire disclosure process and mentality need to shift. AI has fundamentally changed how vulnerabilities are found and exploited. The old model simply doesn't scale anymore.

https://x.com/brian_pak/status/2050255258098766101

Vulnerabilities in the Age of AI

My Predictions – At Least to Mid-Term Years

- More vulnerabilities, more exploits
- Most hacking will be fully autonomous hackbots
- More vulnerabilities, exploited faster
- Higher criticalities
- More zero days as a percentage of total vulns
- More exploited zero days
- Higher % of CVEs used in attacks
- More victims

Vulnerabilities in the Age of AI

Attacks Against the AI You Use

New Mindset Needed

- Many new types of vulnerabilities (ex. Attacks against the AI you use) coming your way!
- You have to update how you think about vulnerabilities
- Vulnerability is the interaction and intersection of many things
- More supply chain attacks
- More interaction attacks

Updating Your Patch Mgmt Program

Defenses

Summary

- All is not lost
- Good guys invented AI
- Good guys use AI more
- Nearly every cybersecurity defense will use AI
- Your patching will need to use AI

Defenses

The “AI Vulnerability Storm”: Building a “Mythos-ready” Security Program

Expedited Strategy Briefing
By the CSA CISO Community, SANS, [un]prompted, the OWASP Gen AI Security Project, and the wider community.

Contact cisos@cloudsecurityalliance.org with any inquiries.

Original release: 12 April, 2026
Last updated: 01 May, 2026

Version 1.0

The latest version of this document can be found here.

cloud security alliance® SANS [un]prompted OWASP GenAI SECURITY PROJECT

Contributing authors include:

Jen Easterly, CEO, RSA Conference, former Director of CISA
Bruce Schneier, Chief of Security Architecture, Inrupt, Fellow and lecturer, Harvard Kennedy School
Chris Inglis, former National Cyber Director, The White House
Rob Joyce, former Cybersecurity Director, NSA
Heather Adkins, CISO, Google
Joshua Saxe, CTO and Co-founder at Security Superintelligence Labs; former AI and Llama Security Lead, Meta
Sounil Yu, CTO, Knostic, former Chief Security Scientist, Bank of America
John N. Stewart, Talons Ventures; former CSTO, Cisco Systems
Katie Moussouris, Founder and CEO, Luta Security
Dave Lewis, Global Advisory CISO, 1Password
Maxim Kovalsky, Managing Director, AI Security CoE, Consortium Networks

Thank you to the 250 CISOs who edited and redlined this live.

Use LLM-based vulnerability discovery and remediation capabilities.

Unlike defensive AI technologies, LLM-based vulnerability discovery capabilities are already mature and can be used to your advantage. Start immediately by asking an agent for a

more than code, from GRC to incident response. Triage and test patches, red team your environment, automate audit data collection, and accelerate security operations overall. Encourage and demand for your team to make use of these agents to accelerate their capabilities, alongside deploying security controls to deploy them responsibly.

Prepare to respond to more incidents.

Run tabletop exercises for multiple simultaneous high-severity incidents occurring within the same week, and have playbooks in place for high-level, critical incidents. Examine how to automate remediation capabilities to the degree possible. Verify and enable mitigating controls such as segmentation, egress filtering, Zero Trust architectures, phishing-resistant MFA, and secrets rotation to limit impact when exploitation occurs. The supply chain will be affected.

Increase focus on the basics.

The basics remain valid and can be prioritized for risks that cannot otherwise be mitigated. Segmentation, patching known vulns, Identity and Access Management, and defense-in-depth/breadth all increase the difficulty for attackers. To lower latent risk, expanding these efforts while there is time is prudent.

We cannot outwork machine-speed threats. Re-prioritize, automate, and prepare for burnout.

The cadence and volume of vulnerability disclosures will exceed anything we have experienced before. Consider how you manage current priorities, and request additional headcount and budget for reserve

https://labs.cloudsecurityalliance.org/mythos-ciso/

Main Conclusion

- The patching cycle will significantly decrease
- You will have hours to a day to patch

- One day, less vulns

Update Your Defenses

- You will need AI to fight AI
- Good bot vs bad bot
- Protect Your Agents to Protect Yourself

Patch Management in the Age of AI

Use Agents to:

- Find vulnerabilities in your environment
- Risk-rank them
- Fix those vulnerabilities according to risk of vuln and asset risk
- Test fixes and resolve problems

- Maybe, one day, AI will create fewer vulnerabilities in the code it writes

Update Your Defenses

Agentic Defenses

- Use AI to find and fix your vulnerabilities
- Always have human-in-the-loop at critical decisions
- Don’t give too much trust too early
- Lots of testing

- Use AI that cares about coding securely

Threat Model Your AI

AI Development Frameworks

- MITRE ATLAS Matrix
- NIST AI Risk Management
- UK Guidelines for secure AI development
- ISO/IEC 42001 international standard
- Google Secure AI Framework
- OWASP GenAI Security project
- OWASP Top 10 for LLM Applications

- Mostly for AI developers, but great for threat modelers
- You should be **THREAT MODELING** your AI development and use!

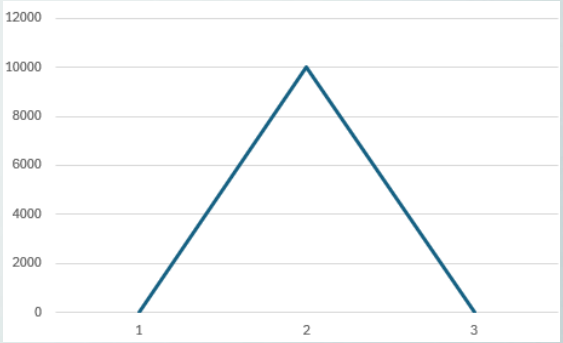
Patch Management in the Age of AI

My Predictions

- Temporary big spike in vulns found and inserted in code
- Temporary big spike in vulns inserted in code from AI
- Huge backlogs of vulns and fixes

Then

- Numbers will decline, perhaps significantly
- But time to patch will shorten permanently



Vulnerability Mgmt in Age of AI

Update Risk Management

- Re-think risk management
  - Faster - meet attacker's speed
  - Is critical risk really critical?
- Re-evaluate tolerance for operational interruption due to bad patches
- Automate security assessments

Vulnerability Mgmt in Age of AI

Faster Patching

- Faster, immediate patching
- You will need AI-enabled patching tools
- “Hot patching”, “micro-patching”??
- Enterprise-Wide Auto-patching??
- Better patching recovery??

Vulnerability Mgmt in Age of AI

Continuous Vuln Scanning

- You're always scanning for vulns
- Don't let the attacker be the only one
- Faster and more aggressive on high-risk assets and networks

Vulnerability Mgmt in Age of AI

Other Things Beyond Patching

Don’t rely only on patching to stop vuln attacks

- Aggressive Logging (and AI to figure out important stuff)
- Aggressive intrusion detection
- Aggressive incident response
- THREAT MODEL!!
- Other protections (e.g., “inline patching”, application-level firewalls, etc.)

Vulnerability Mgmt in Age of AI

Other Things Beyond Patching

Don't rely only on patching to stop vuln attacks

- Defense-in-Depth
- PHISHING-RESISTANT MFA
- Increased focus on 3rd Party Risk Management, especially Supply chain, especially AI integrations
- Segmentation
- Egress filtering
- Deception technology (honeypots)

Vulnerability Mgmt in Age of AI

Conclusion

- Not the end of the world
- But now is NOT the time to wait and see
- Need more AI…on defender's side
- Figure out how to defend against attacks against the AI that you use
- Continuous vuln scanning
- Faster patching
- More more than patching

Questions?

Roger A. Grimes– CISO Advisor, KnowBe4
e: rogerg@knowbe4.com
LinkedIn: https://www.linkedin.com/in/rogeragrimes/ (over 42K followers)
X/Twitter: @RogerAGrimes
Mastodon: https://infosec.exchange/@rogeragrimes
YouTube: @CyberSecWTFRants
Bluesky: rogeragrimes@bsky.social
Reddit: u/rogeragrimes