knowbe4

AI Attacks

What You Need to Know

Roger Grimes
CISO Advisor

Roger A. Grimes

CISO Advisor, KnowBe4

39 Years+ in Cybersecurity, 8+ years at KnowBe4

QUICK BIO

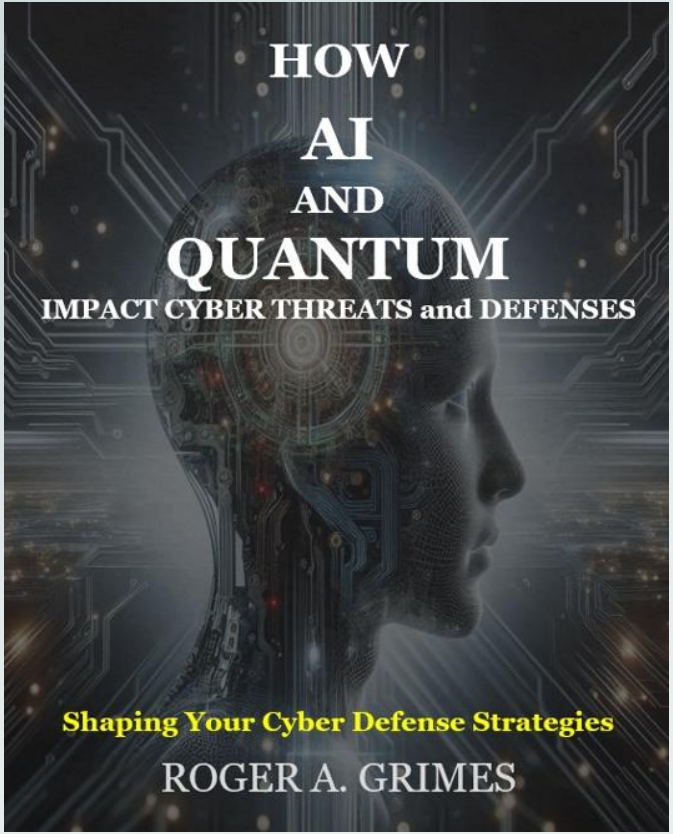
- Started fighting computer viruses in 1987 with John McAfee
- Previously worked for Microsoft, Foundstone, and McAfee
- Previously VP of IT, Network Manager, 20 years pen testing, cybersecurity instructor
- Consultant to world's largest to smallest companies plus militaries for decades
- Career goal - 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

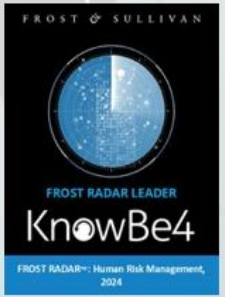
Global Sales, Courseware Development, Customer Success, and Technical Support teams worldwide

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

- Traditional Attacks
- What is AI?
- AI Deepfake Attacks Over Time
- Defenses

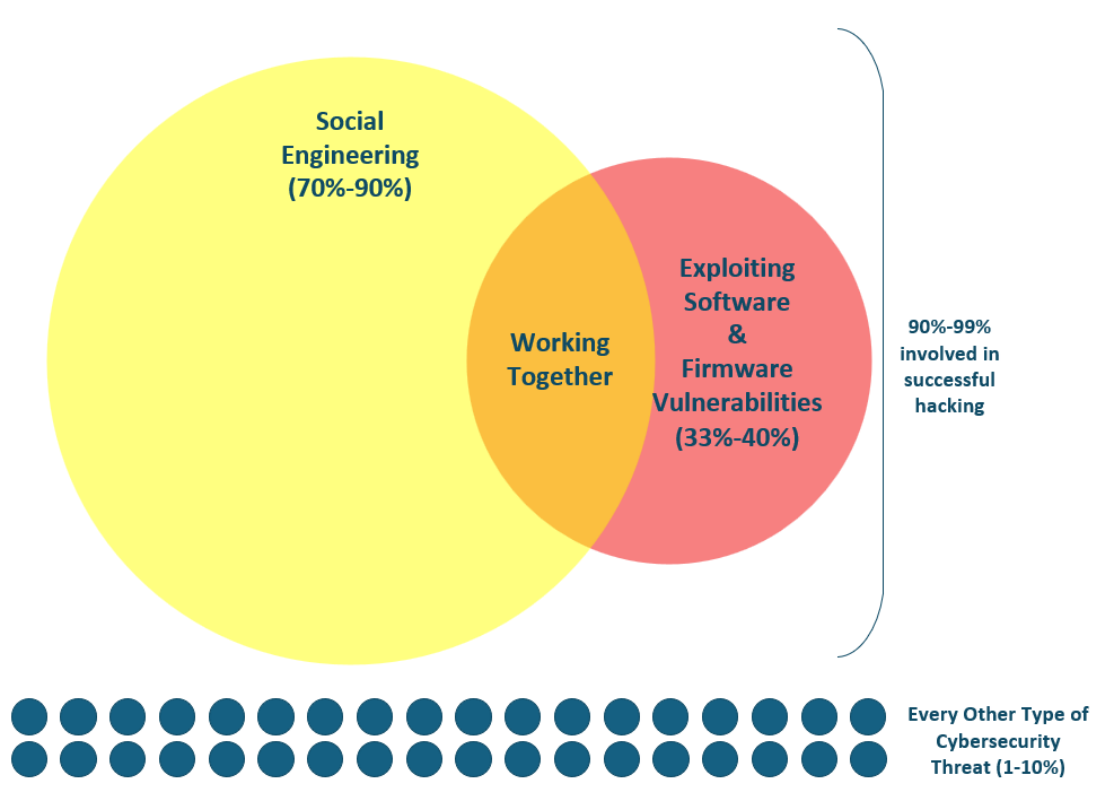
Agenda

- Traditional Attacks
- What is AI?
- AI Deepfake Attacks Over Time
- Defenses

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



This will likely be how attackers use AI as well

Agenda

- Traditional Attacks
- What is AI?
- AI Deepfake Attacks Over Time
- Defenses

What Really Is AI?

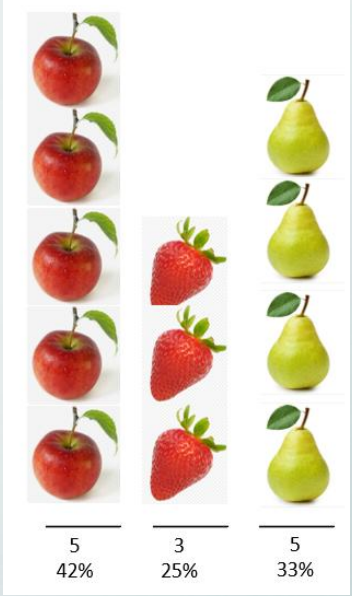
Traditionally Understood Definition:

A system or service that is able to perform tasks simulating “human intelligence” when learning, reasoning, and decision-making.

What Really Is AI?

What AI Really Is:

- 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)
- Can communicate using human language



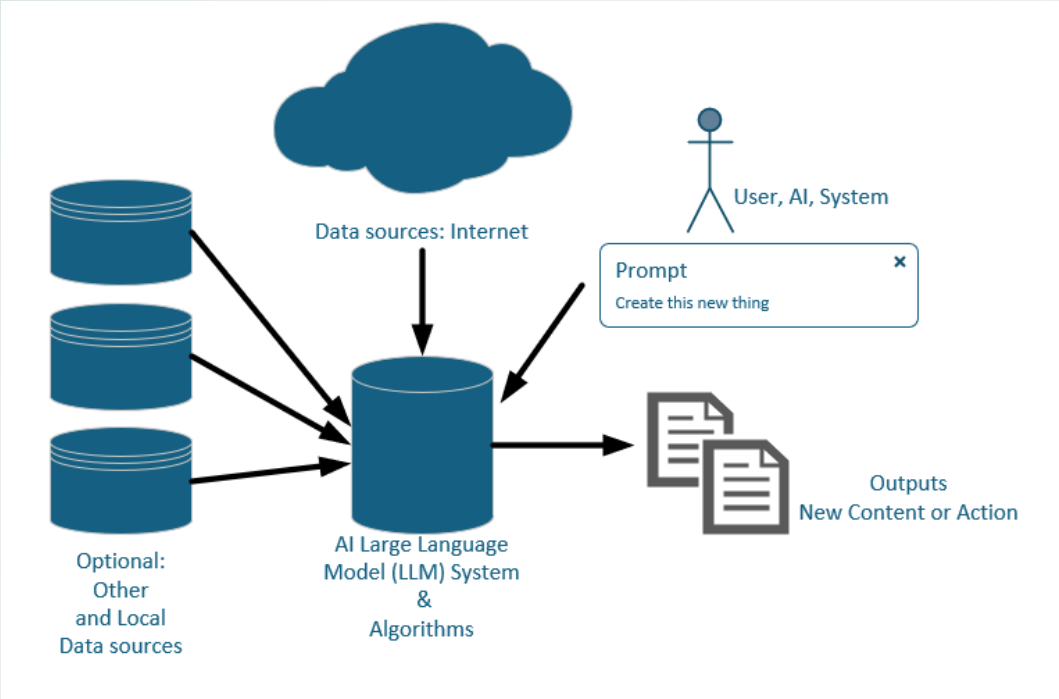
What Really Is AI?

What AI Really Isn’t:

- It’s not human
- It’s software

What Is AI?

Large Language Model (LLM)



What Is Agentic AI?

Generative AI

General Definition

- **A type of LLM AI that creates new, original content such as text, images, audio, video, and other media**
- Some definitions of generative AI include any AI that creates brand new anything, such as protein folding
- But most public definitions limit generative AI to media-type content
- Currently, lots of mistakes “hallucinations”, but getting better all the time

https://www.linkedin.com/pulse/what-agentic-ai-roger-grimes-420ve

Malicious AI Threats

AI-Enabled Deepfakes

- “Deepfake” is a very realistic, but fake picture, video, or audio recording of a real or fake person
- All someone needs is a picture of that person and 6-30 seconds of audio
- Can be used to create very realistic phishing attacks
- We are already seeing very good AI “deepfakes” used to exploit people

What Is Agentic AI?

Deepfakes

General Definition

- **Fake/"synthetic" media content created by generative AI**
- Thousands of sites

Created on Hedra.ai

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

What Is Agentic AI?

Autonomous Agentic AI

General Definition

- **Some additional level of self-control and decisions (i.e., autonomy) over traditional software**
- **Ranges from little to no autonomy, quasi-autonomy, to full autonomy**
- High autonomy will have growing pains
- High autonomy will cause unexpected problems
- No way to completely test in all scenarios
- You are handing a lot of trust to a high autonomous AI
- Potentially scary – Can update its own goals
  - **Terminator Skynet becomes self-aware on 2:14 a.m., EDT, on August 29, 1997!!**

Agenda

- Traditional Attacks
- What is AI?
- AI Deepfake Attacks Over Time
- Defenses

Malicious LLMs – So 2023

Dear [Employee Name],

I hope this email finds you in good health and high spirits. I am writing to you today with a surprise that I believe will lift your spirits even higher.

As you may be aware, our company has been experiencing financial success of late. I am pleased to announce that this success has allowed us to grant our hard-working employees a pay raise. You, [Employee Name], are one of those employees.

Attached to this email, you will find a document detailing the specifics of your raise. Please review it at your earliest convenience and do not hesitate to reach out to me with any questions.

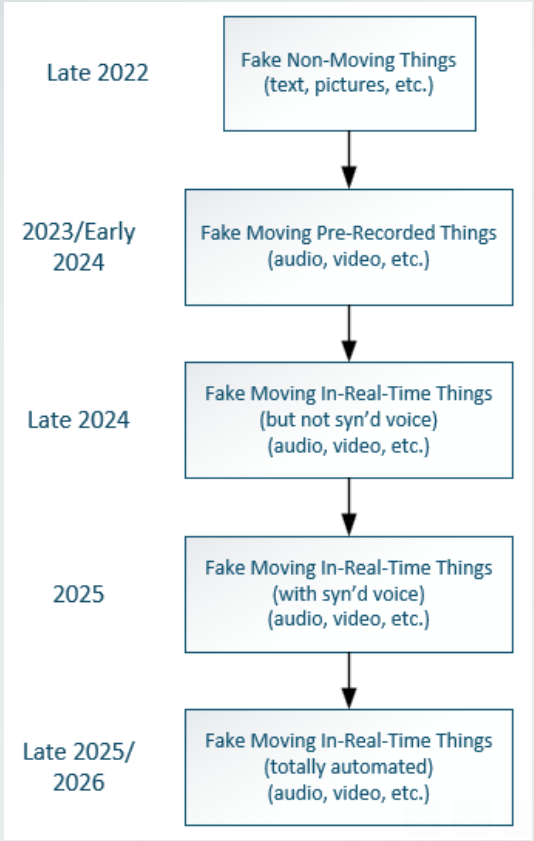
Your hard work and dedication to our company have not gone unnoticed, and I am thrilled to be able to recognize your contributions in this way.

Once again, congratulations on your pay raise. Keep up the great work.

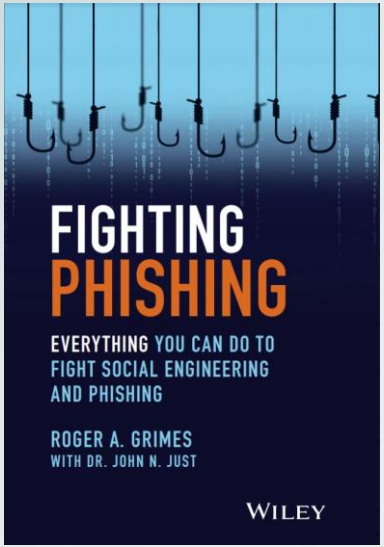
Best regards,

[Your Name]

AI-Enabled Deepfakes Maturation



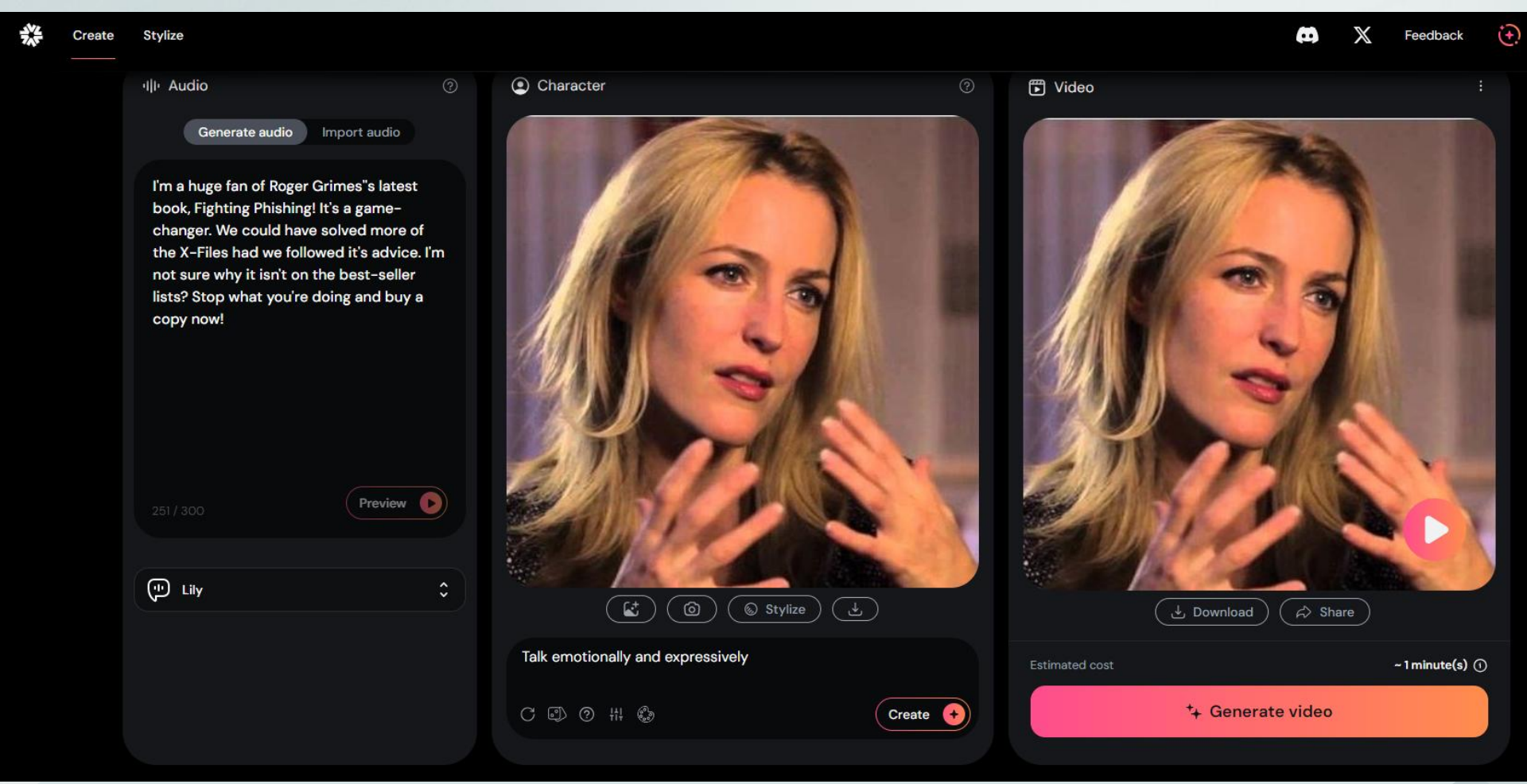
Apparently, actress Gillian Anderson loves one of my books



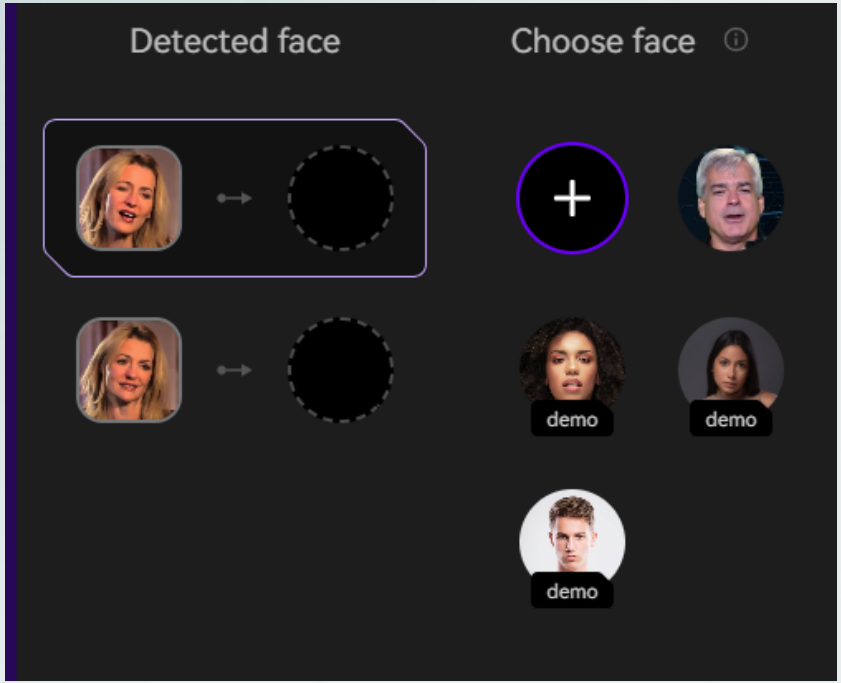
I did this in the first 30 seconds during my first time on an AI site, my first time ever trying to make a AI-created deepfake video

Imagine what anyone could do in 5 minutes?

https://www.hedra.com/



Deepfakes



Swapped Faces

www.akool.com

Deepfakes

Face Swap
&
AI-Generated Voice

Uploaded Face
Swapped Picture
Uploaded Fake Audio
Took 30 seconds

https://www.hedra.com/

Deepfakes Are Getting Realistic

Examples

AI/Deepfakes

Just making long-game phishing easier to do

Hong Kong police said at a press conference Friday that the employee at the unnamed firm's Hong Kong branch initially suspected phishing when he received an email last month purporting to be from the company's UK-based chief financial officer, CNN reported.

However, after attending a video conference and seeing convincing deepfakes of the CFO and other colleagues, the employee believed the request to carry out a secret transaction was legitimate.

https://www.scmagazine.com/news/deepfake-video-conference-convinces-employee-to-send-25m-to-scammers

Deepfake video conference convinces employee to send $25M to scammers

Laura French February 5, 2024

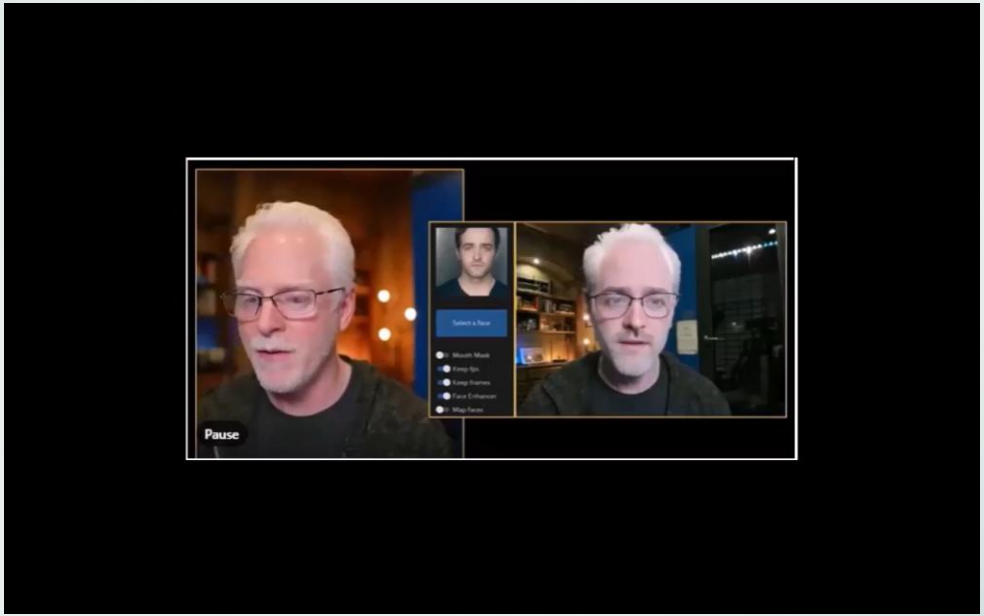
An employee was tricked into sending $35 million to scammers after seeing colleagues on a video call that turned out to be deepfaked, police say. (Credit: Adobe Stock)

A deepfake phishing scam cost a multinational company more than $25 million after an employee was fooled by digital imitations of his colleagues on a conference call.

Real-Time AI-Enabled Deepfakes

Fake Real-Time Video

Hackers can create real-time AI fakes that immediately duplicate what they are saying and doing



https://www.linkedin.com/pulse/real-time-ai-agents-already-here-roger-grimes-wphge

Real-Time AI-Enabled Deepfakes

Fake Real-Time Video

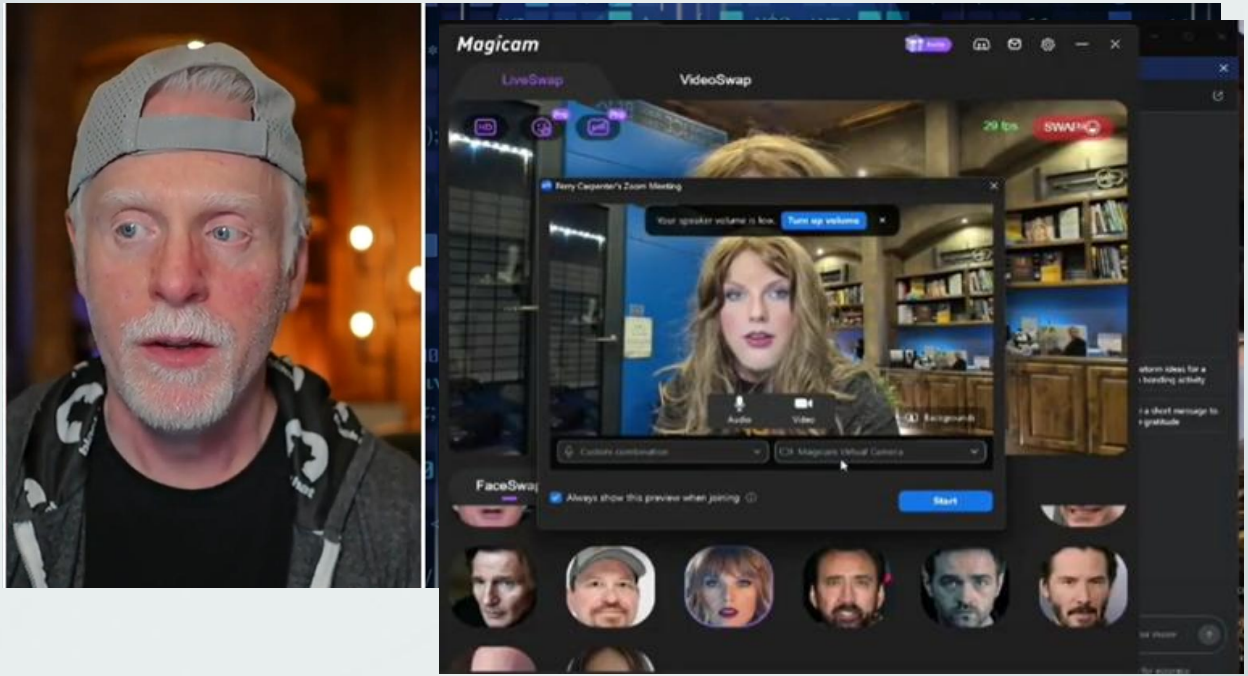
Pick Your Deepfake Identity With a Mouse Click

https://www.linkedin.com/posts/perrycarpenter_cybersecurity-securityawareness-ai-activity-7318408334393384960-_vQD

Real-Time AI-Enabled Deepfakes

Pick Your Deepfake Identity With a Mouse Click

Fake Real-Time Video

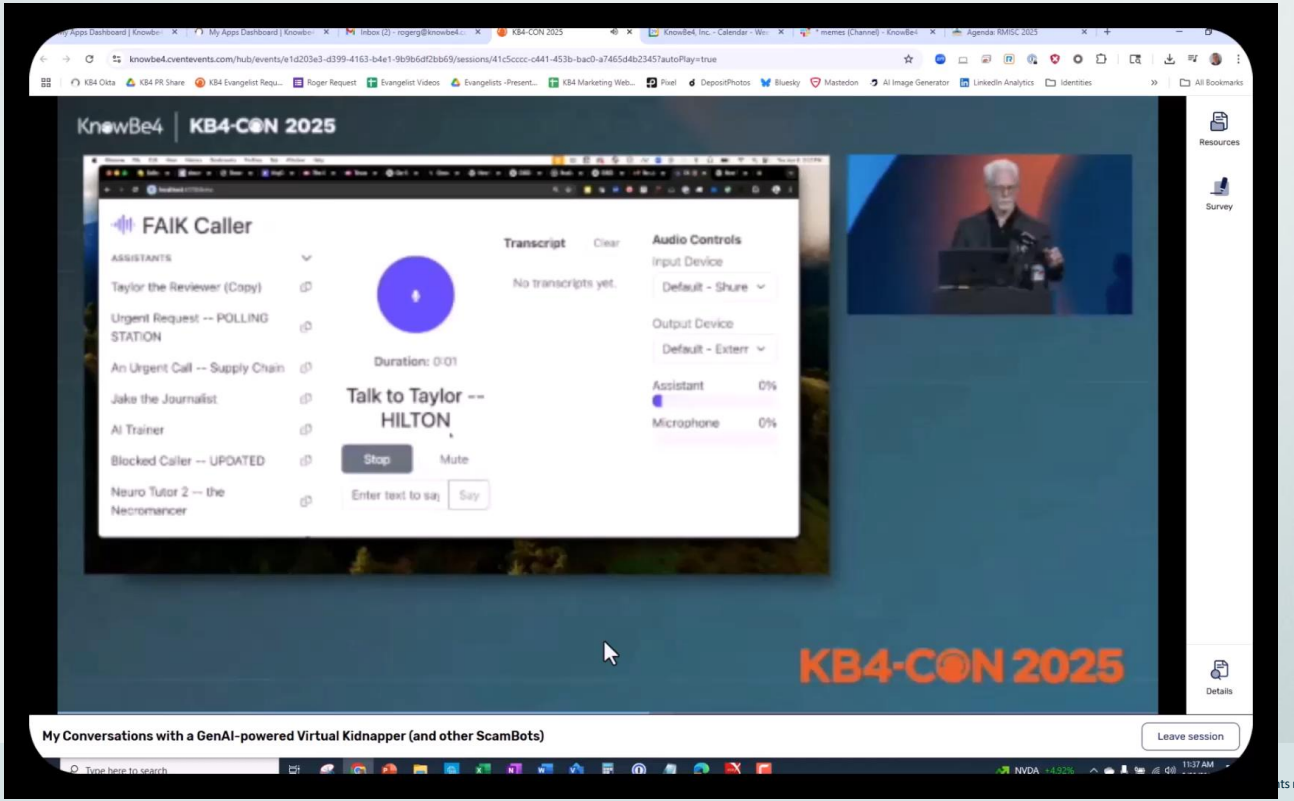


https://www.linkedin.com/posts/perrycarpenter_cybersecurity-securityawareness-ai-activity-7318408334393384960-_vQD

Real-Time AI-Enabled Deepfakes

Fake Real-Time Video

AI-Enabled LLM Does Entire Real-Time Conversation



Real-Time AI-Enabled Deepfakes

Fake Real-Time Video

Deepfaking Zoom Calls

https://www.linkedin.com/posts/perrycarpenter_cybersecurity-securityawareness-ai-activity-7318408334393384960-_vQD

It’s Automated Now

Reused Stolen Biometrics

Hackers and Malware

Face Off: Group-IB identifies first iOS trojan stealing facial recognition data

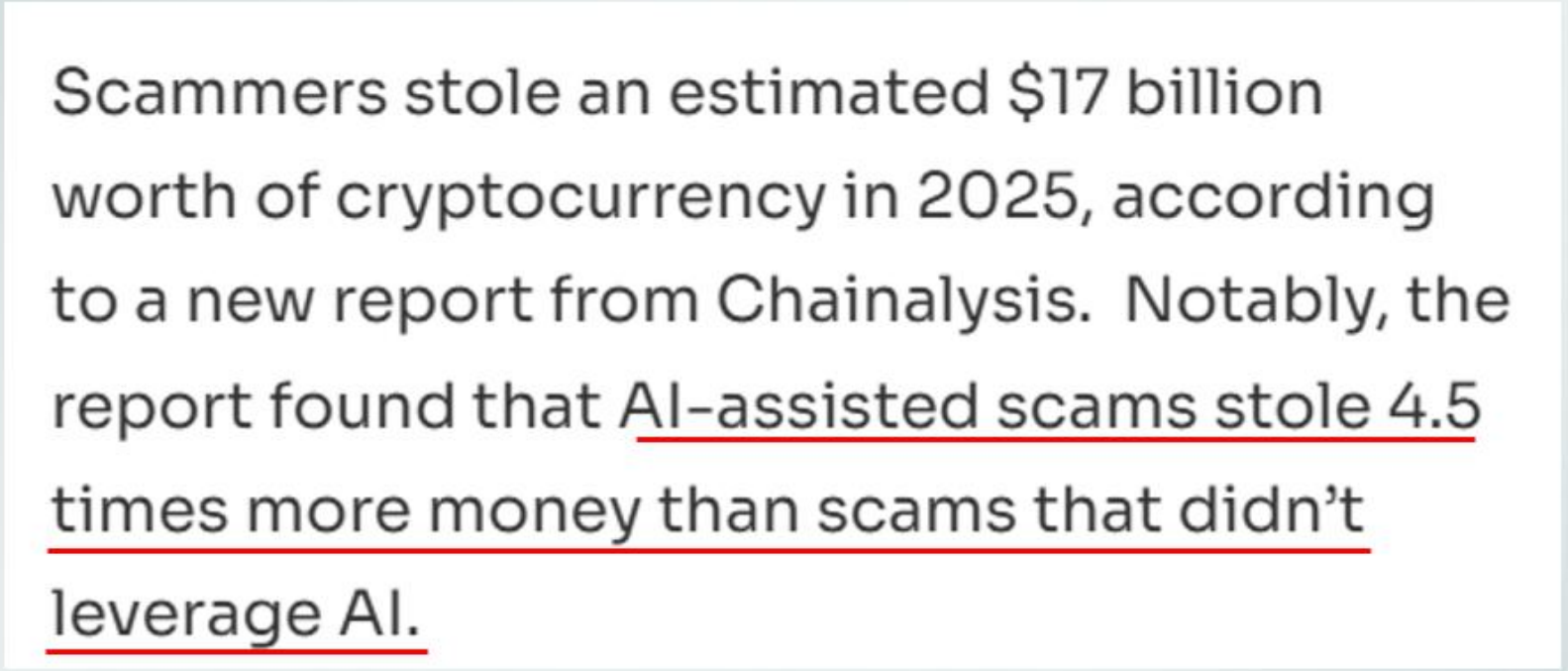
Group-IB uncovers the first iOS Trojan harvesting facial recognition data used for unauthorized access to bank accounts. The GoldDigger family grows

February 15, 2024 · 47 min to read · Malware Analysis

https://www.linkedin.com/pulse/game-changer-biometric-stealing-malware-roger-grimes-ikaze

https://www.group-ib.com/blog/goldfactory-ios-trojan/

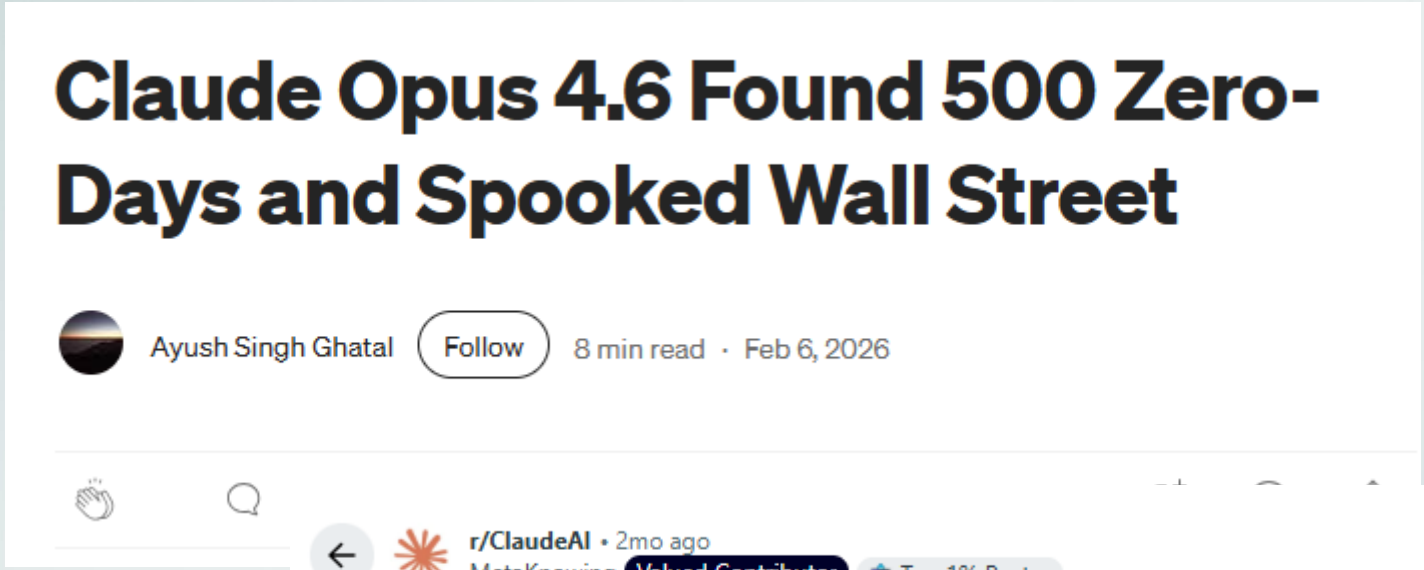
AI-Enabled Attacks Are More Profitable



https://blog.knowbe4.com/report-scammers-stole-17-billion-worth-of-crypto-last-year

Vulnerability Hacking Bots

AI Vulnerability Hacking Bots



AI Attack Example

AI-Enabled Hackbots

- Find and quickly exploit every newly announced vulnerability
- Find and exploit new 0-days

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

We have identified thousands of additional high- and critical-severity vulnerabilities that we are working on responsibly disclosing to open source maintainers and closed source vendors.

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 chatbo 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.

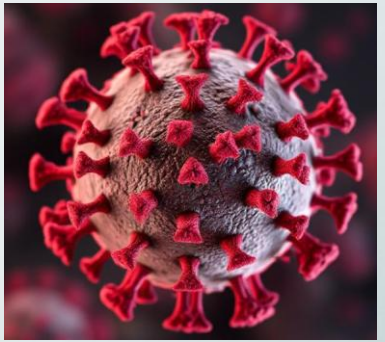
The chatbot then analyzed the companies' hacked financial documents to help determine a realistic amount of bitcoin to demand in exchange for the hacker's promise not to publish that material. It also wrote suggested extortion emails.

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

What Is Agentic AI Malware?

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

Language Models Can Autonomously Hack and Self-Replicate

2026-05-07 | Alena Air, Reworr, Nikolaj Kotov, Dmitrii Volkov, John Steidley, Jeffrey Ladish

Highly Autonomous Malware

Not Here Yet

But

It Will Be Soon

It’s Starting - Agentic AI Ransomware

ESET discovers PromptLock, the first AI-powered ransomware

BRATISLAVA — August 27, 2025 — ESET researchers have uncovered a new type of ransomware that leverages generative artificial intelligence (GenAI) to execute attacks. Named PromptLock, the malware runs a locally accessible AI language model to generate malicious scripts in real time. During infection, the AI autonomously decides which files to search, copy, or encrypt — marking a potential turning point in how cybercriminals operate.

https://www.eset.com/us/about/newsroom/research/eset-discovers-promptlock-the-first-ai-powered-ransomware

Fully Agentic AI Attacks

Google Threat Intelligence Group Updates on AI Attacks, May 11th 2026

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

Autonomous Malware Operations: AI-enabled malware, such as PROMPTSPY, signal a shift toward autonomous attack orchestration, where models interpret system states to dynamically generate commands and manipulate victim environments. Our analysis of this

AI-Augmented Research and IO: Adversaries continue to leverage AI as a high speed research assistant for attack lifecycle support, while shifting toward agentic workflows to operationalize autonomous attack frameworks. In information operations (IO) campaigns,

https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access

Agenda

- Traditional Attacks
- What is AI?
- AI Deepfake Attacks Over Time
- Defenses

Defenses

Summary

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

Malicious Agentic AI Deepfake Threats

Luckily

- Not all is lost
- We have AI threat-detecting and threat-hunting on our side to fight back

It will be:

- Good AI-enabled bot versus bad AI-enabled bot
- Best algorithm wins

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

Defenses

Overview

- Education, education, education

Best Defenses

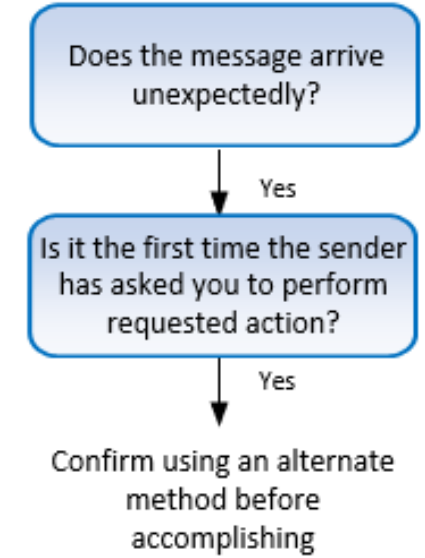
Teach This!

How

To

Spot

Phishing



https://blog.knowbe4.com/teach-two-things-to-decrease-phishing-attack-success

Best AI Defenses

To Prevent Malicious AI Deepfakes Being Used Against You

- Establish a “safe word” with your family
  - Potentially for business as well
- When in doubt, ask the other person a personal question that only the real person would know (ex. “What’s your favorite book, again?”)
- Warn them about the type of Deepfake scams that target family members
  - Fake kidnappings
  - Fake accidents
  - Fake legal problems

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