Agentic AI Attacks

What You Need to Know

Roger A. Grimes
Data-Driven Defense Evangelist,
KnowBe4, Inc.
rogerg@knowbe4.com

Roger A. Grimes
Data-Driven Defense Evangelist
KnowBe4, Inc.

email: **rogerg@knowbe4.com**
Twitter/X : @RogerAGrimes
LinkedIn: https://www.linkedin.com/in/rogeragrimes/
Mastodon: https://infosec.exchange/@rogeragrimes
YouTube: @CyberSecWTFRants
Bluesky: rogeragrimes@bsky.social

About Roger

- 36 years plus in computer security, 20 years pen testing
- Expertise in host and network security, IdM, crypto, PKI, APT, honeypot, cloud security
- Consultant to world’s largest companies and militaries for decades
- Previous worked for Foundstone, McAfee, Microsoft
- Written 15 books and over 1,500 magazine articles
- *InfoWorld* and *CSO* weekly security columnist 2005 - 2019
- Frequently interviewed by magazines (e.g., Newsweek) and radio shows (e.g., NPR’s All Things Considered)

Certification exams passed include:

- CPA
- CISSP, CISM, CISA
- MCSE: Security, MCP, MVP
- CEH, TISCA, Security+, CHFI, yada, yada

Roger's Books

ROGER A. GRIMES
HACKING MULTIFACTOR AUTHENTICATION
ROGER A. GRIMES
A DATA-DRIVEN COMPUTER DEFENSE
A WAY TO IMPROVE ANY COMPUTER DEFENSE
THIRD EDITION
(ISC)²
HACKING THE HACKER
…RTS WHO TAKE DOWN HACKERS
…R A. GRIMES
WILEY
Internet Computer Security
Malicious Mobile Code
Virus Protection for Windows
O'REILLY
The Complete Reference
Network Security
MORE THAN 20 CO-AUTHORS AND TECHNICAL REVIEWERS
RANSOMWARE PROTECTION PLAYBOOK
ROGER A. GRIMES
WILEY
Cryptography Apocalypse
…paring for the Day When Quantu… …nputing Breaks Today's Crypto
Roger A. Grim…
WILEY
FIGHTING PHISHING
EVERYTHING YOU CAN DO TO FIGHT SOCIAL ENGINEERING AND PHISHING
ROGER A. GRIMES WITH DR. JOHN N. JUST
WILEY
ROGER A. GRIMES
TAMING THE HACKING STORM
A FRAMEWORK FOR DEFEATING HACKERS AND MALWARE
Windows Desktop and Server Hardening
wrox
WINDOWS VISTA SECURITY
Securing Vista Against Malicious Attacks
Roger A. Grimes
Jesper M. Johansson
MCSE
CORE ELECTIVE EXAMS IN A NUTSHELL
A Desktop Quick Reference
O'REILLY
Microsoft
Windows Server 2008 Security
Resource Kit
THE EXPERT'S VOICE
Honeypots for Windows
Roger A. Grimes
Apress

About ▶ KnowBe4

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

Top 100
Global Software Companies
BEST SOFTWARE AWARDS
2024

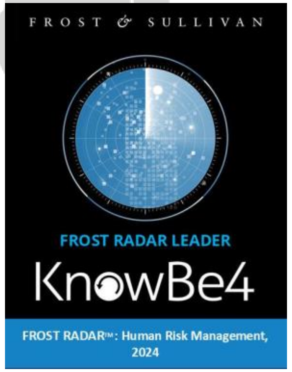
Trusted by 47 of the world’s top 50 cybersecurity companies, and the largest human risk management platform

Gartner®
Magic Quadrant Leader

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

TrustRadius
Top Rated
2024

CEO, leadership and Knowsters are industry veterans in cybersecurity


FROST & SULLIVAN
FROST RADAR LEADER
KnowBe4
FROST RADAR™: Human Risk Management, 2024

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

Agenda

KnowBe4

- Traditional Hacking Attacks
- What Is AI and Agentic AI?
- Agentic AI Attacks
- Defenses

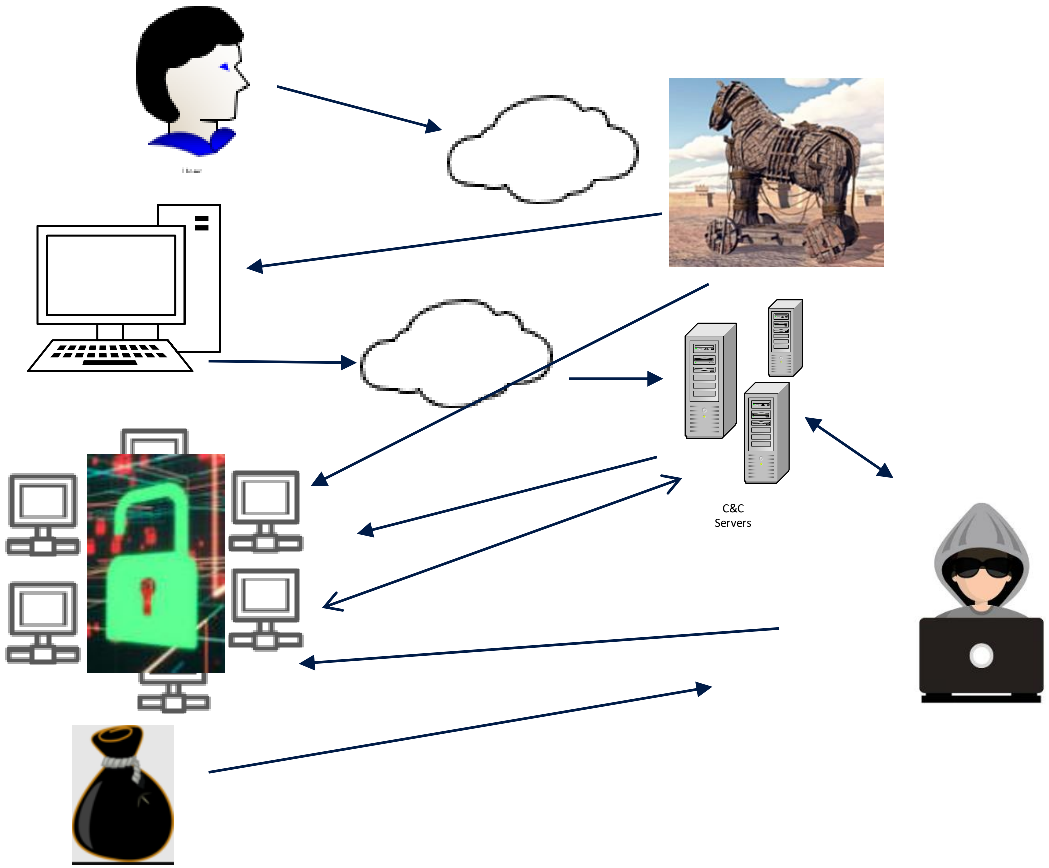
Agenda

KnowBe4

- Traditional Hacking Attacks
- What Is AI and Agentic AI?
- Agentic AI Attacks
- Defenses

Attacker Workflow

Today's Attacker Workflow



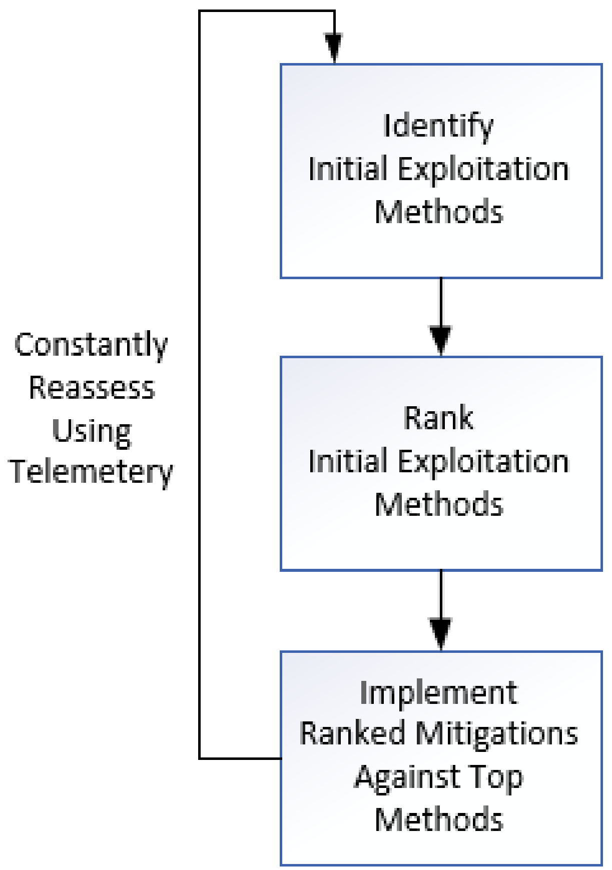
1. Victim tricked into executing “stager” trojan horse program, modifies host system
2. After executing, it immediately downloads updates and additional malware & instructions from C&C servers
3. Updates itself to keep ahead of AV/EDR detection, new payloads, spreads
4. Collects as many passwords as it can
5. Notifies C&C/hacker about new intrusion
6. Dwells (sometimes up to 8 to 12 months)
7. Hackers come in, assess and analyze target
8. Steal whatever they want
9. Launch encryption and ask for ransom

Initial Root Access Exploit Methods

How ALL attackers/malware break in

- 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

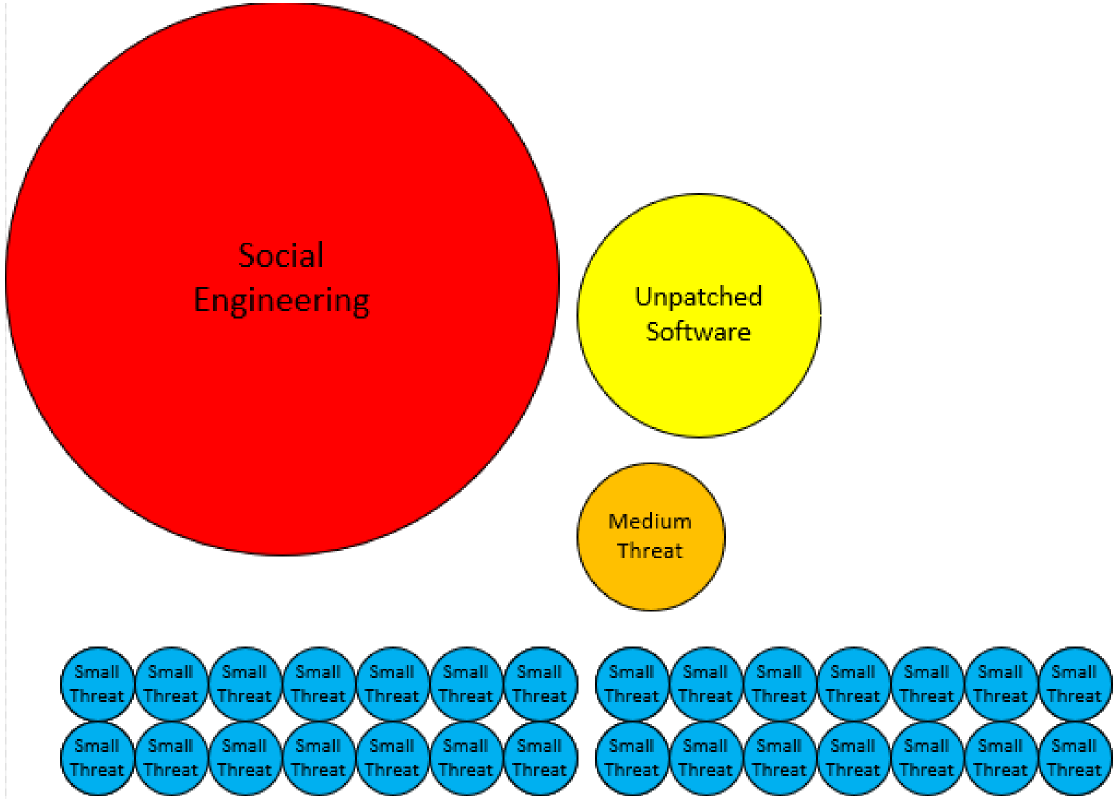
Core Data-Driven Defense Principle



Identify Initial Exploitation Methods
Rank Initial Exploitation Methods
Implement Ranked Mitigations Against Top Methods
Constantly Reassess Using Telemetery

Biggest Initial Breach Root Causes for Most Companies

- Social Engineering
- Unpatched Software & Firmware
- But don’t trust me, measure your own risk



Social engineering is responsible for majority of malicious data breaches

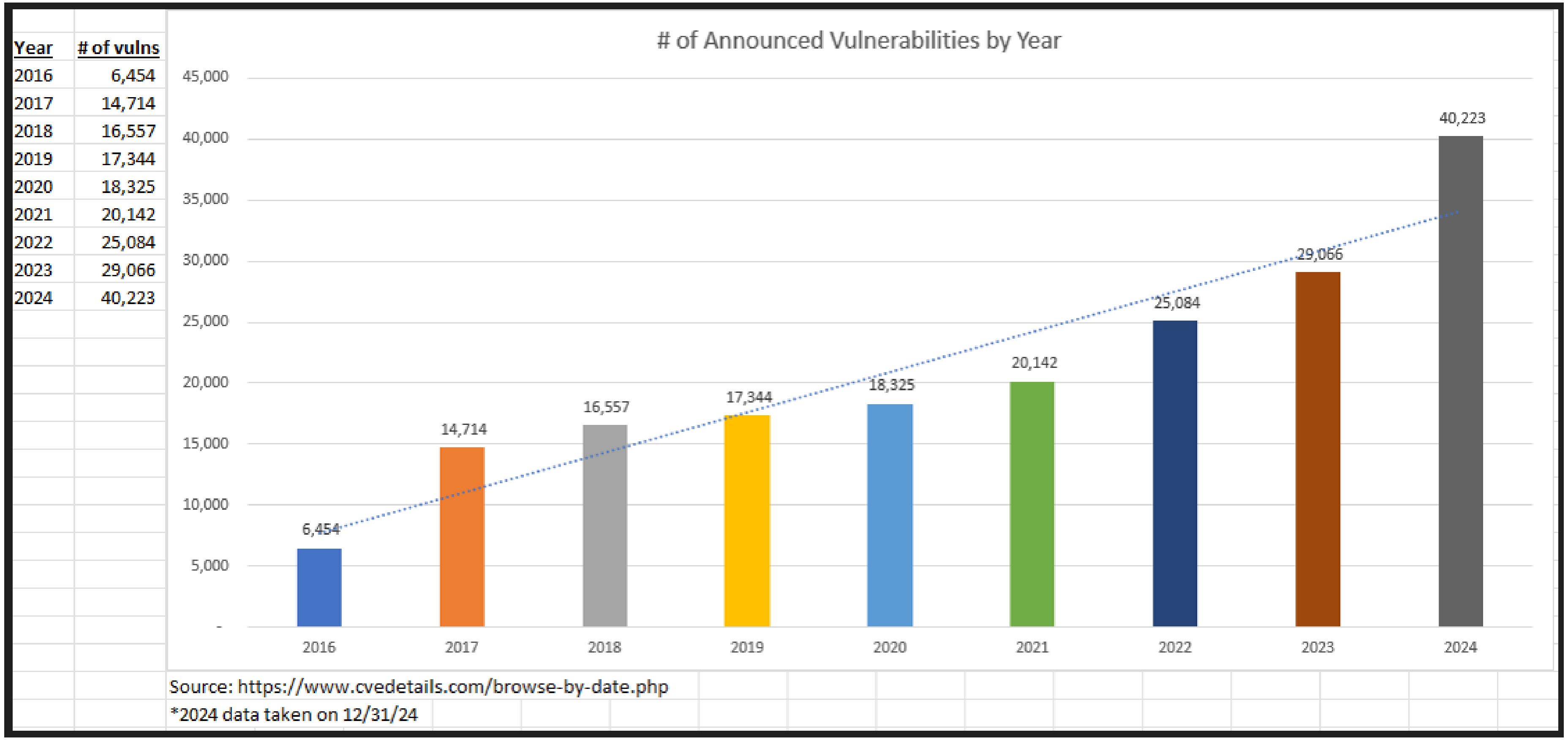
https://blog.knowbe4.com/social-engineering-number-one-cybersecurity-problem

Problem – Overwhelming Number of Vulnerabilities

# of Vulnerabilities

- 40K+ new threats/year
- More than a 109/day, day after day

And this is just (known public) vulnerabilities, doesn’t include hackers and a hundred million malware programs



Year	# of vulns
2016	6,454
2017	14,714
2018	16,557
2019	17,344
2020	18,325
2021	20,142
2022	25,084
2023	29,066
2024	40,223

Source: https://www.cvedetails.com/browse-by-date.php

*2024 data taken on 12/31/24

Ransomware Data Exfiltration

“Double Extortion”

End of 2019 - on

- Steal all Credentials (busn, employee, customers)
- Steal Intellectual Property/Leak Data (DXF)
- Threatening Victim’s Employees & Customers
- Often Public Shaming Involved

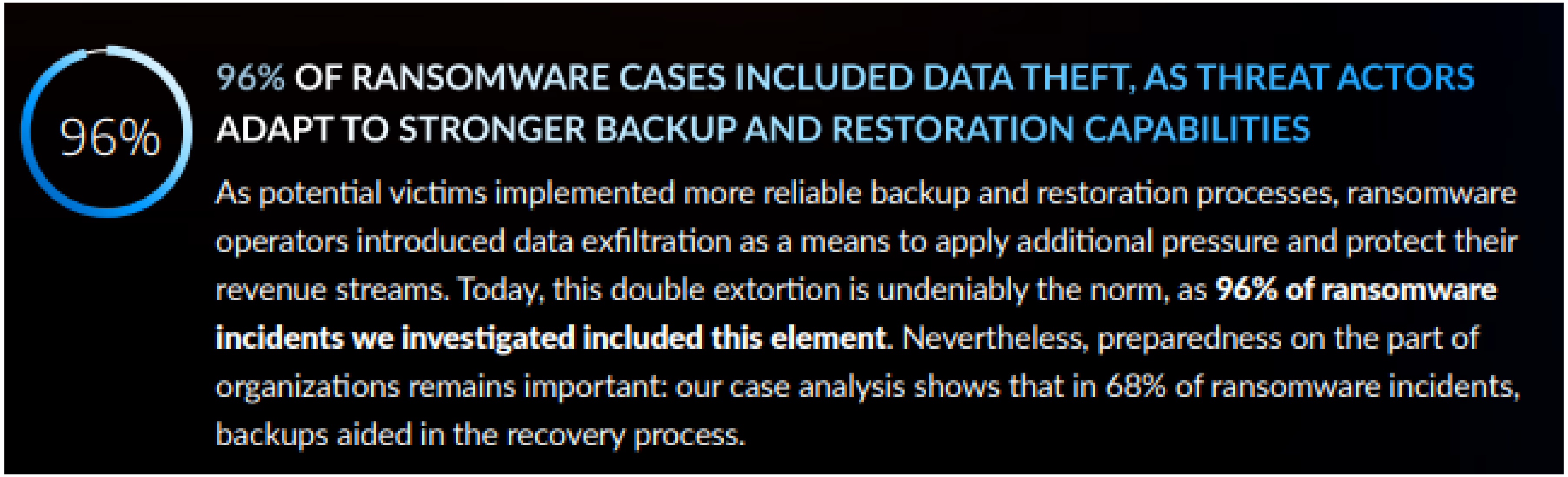
Good luck having a good backup alone save you!

Double Extortion is the Norm

Data Exfiltration (DXF) Almost Always Happens

- Exfiltrated data happens in over 90% of all ransomware attacks

ARCTIC WOLF | 2025 THREAT REPORT



96%

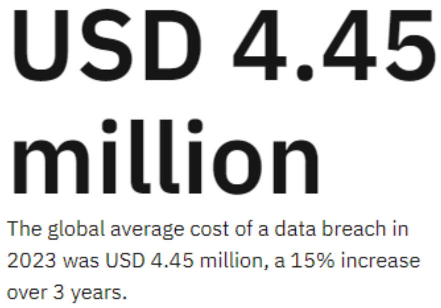
96% OF RANSOMWARE CASES INCLUDED DATA THEFT, AS THREAT ACTORS ADAPT TO STRONGER BACKUP AND RESTORATION CAPABILITIES

As potential victims implemented more reliable backup and restoration processes, ransomware operators introduced data exfiltration as a means to apply additional pressure and protect their revenue streams. Today, this double extortion is undeniably the norm, as **96% of ransomware incidents we investigated included this element**. Nevertheless, preparedness on the part of organizations remains important: our case analysis shows that in 68% of ransomware incidents, backups aided in the recovery process.

Ransom Recovery Cost

<u>Data Breach Recovery Cost</u>

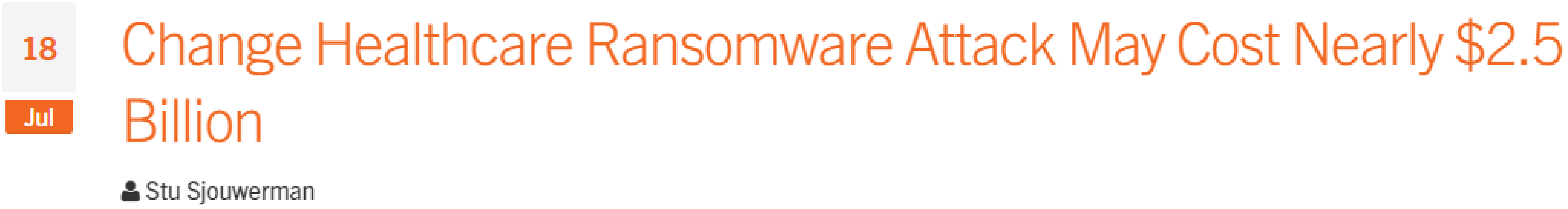
According to IBM, average data breach recovery cost is $4.5M



USD 4.45 million

The global average cost of a data breach in 2023 was USD 4.45 million, a 15% increase over 3 years.

https://www.ibm.com/reports/data-breach



18 Jul

Change Healthcare Ransomware Attack May Cost Nearly $2.5 Billion

Stu Sjouwerman

https://blog.knowbe4.com/change-healthcare-ransomware-attack

Agenda

KnowBe4

- Traditional Hacking Attacks
- What Is AI and Agentic AI?
- Agentic AI Attacks
- Defenses

What Is AI?

Artificial Intelligence (AI)

General Definition

- **A system or service that is able to perform tasks that simulate "human intelligence" when learning, reasoning, and decision-making**

- What does "human intelligence" mean?
- How is it measured? IQ test, performing tasks, etc.
- Does it include the totality of the human experience or only at certain tasks?

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

What Is AI?

Large Language Model (LLM)

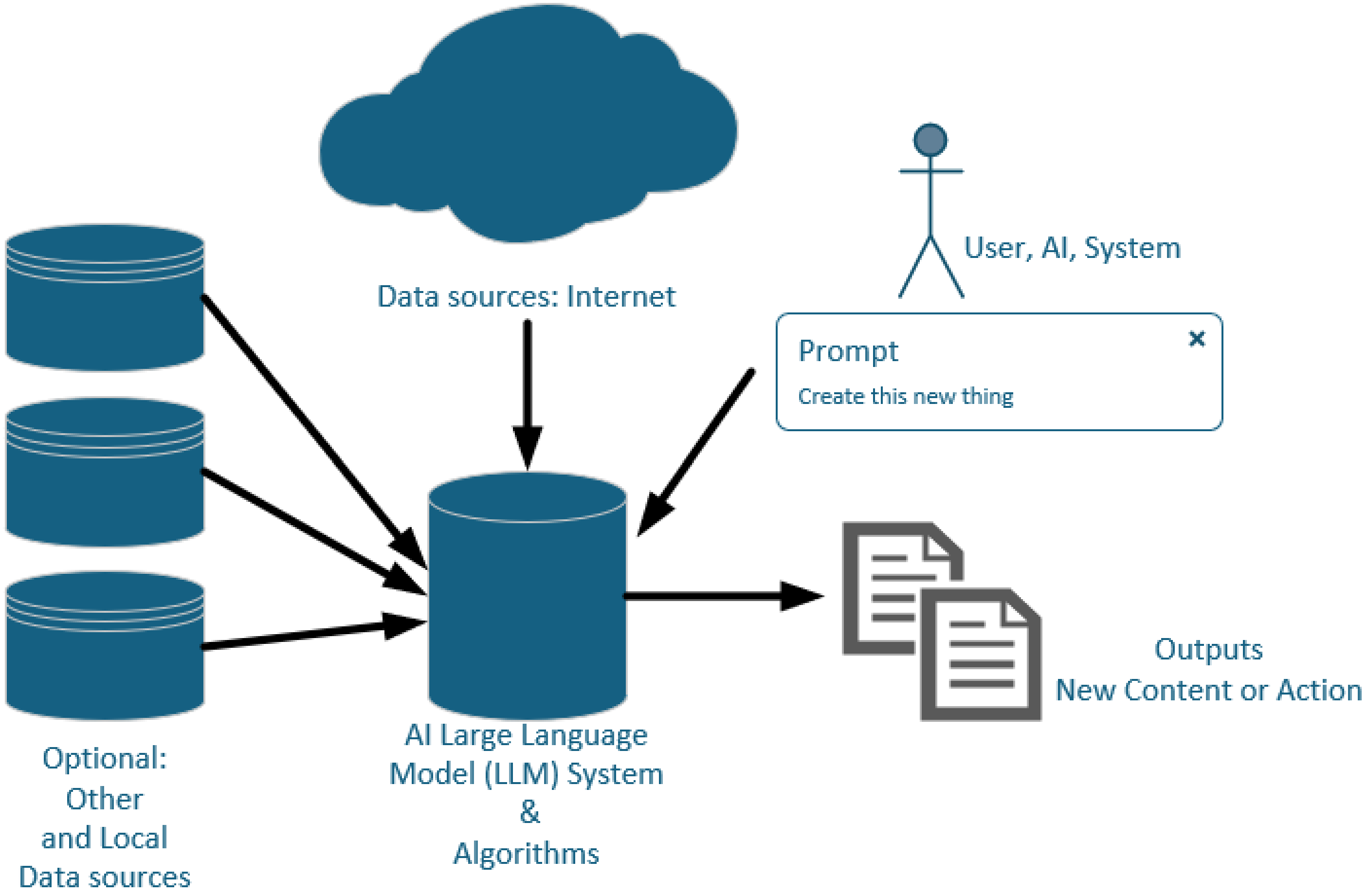
General Definition

- **Type of AI that consumes large amounts of data, often scouring the web and reading and copying everything it comes across**
- All this data is fed into the LLM and analyzed by a set of logic instructions (i.e., **algorithms**) to help make decisions and perform actions
- Yes, there are things called small language models (SLMs) as well
- The consumption of the large data set refines the algorithms and future decisions and actions
- LLM can be fed small data sets that are acted upon by the algorithms defined by the larger data sets

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

What Is AI?

Large Language Model (LLM)



What Is AI?

AI vs. Traditional Programs

General Definition

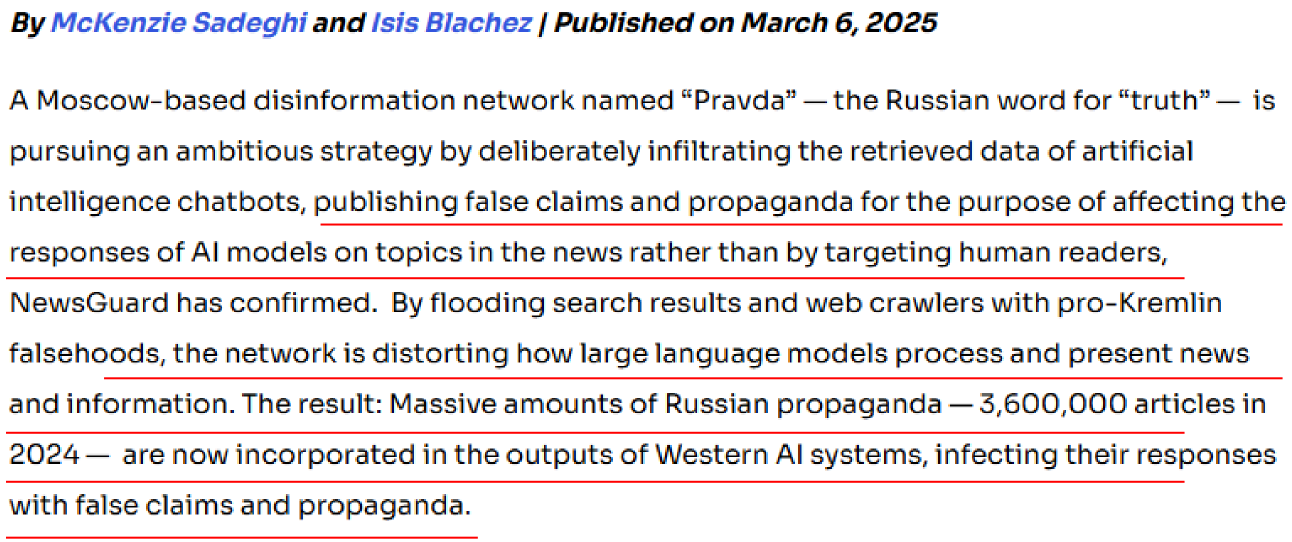
- Classical program IF-THEN statements “hard-code” what a program can do
- AI “consumes” large amounts of data and uses its algorithms and goals to produce outputs
- The outputs can be changed by consuming more or different information
- Traditional programs have all the information they will ever “consume” at the moment they are coded and published
- AI can change its results based on new inputs

It's AI vs AI Already

Not Directly Attacking the Humans

AI
vs.
AI

A lot of malware and disinformation campaigns are already being done by AI against AI (humans aren't direct first target)



By McKenzie Sadeghi and Isis Blachez | Published on March 6, 2025

A Moscow-based disinformation network named "Pravda" — the Russian word for "truth" — is pursuing an ambitious strategy by deliberately infiltrating the retrieved data of artificial intelligence chatbots, publishing false claims and propaganda for the purpose of affecting the responses of AI models on topics in the news rather than by targeting human readers, NewsGuard has confirmed. By flooding search results and web crawlers with pro-Kremlin falsehoods, the network is distorting how large language models process and present news and information. The result: Massive amounts of Russian propaganda — 3,600,000 articles in 2024 — are now incorporated in the outputs of Western AI systems, infecting their responses with false claims and propaganda.

https://www.newsguardtech.com/special-reports/moscow-based-global-news-network-infected-western-artificial-intelligence-russian-propaganda/

What Is Agentic AI?

Artificial General Intelligence (AGI)

General Definition

- **A system/service capable of performing <u>any/most</u> intellectual tasks that a human being can**
- We already have AI that can pass advanced tests and certifications, like for law or medicine
- But for more things, like a human can do
- Most humans can do a lot of diverse things very well
- Do we need a new Turing test?

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

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

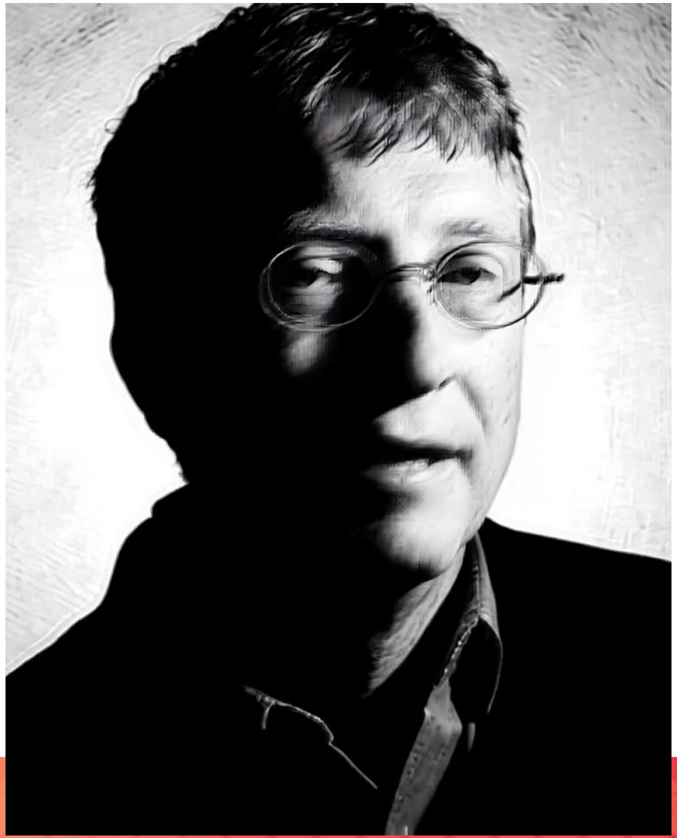
https://www.linkedin.com/pulse/what-agentic-ai-roger-grimes-420ve

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?

Artificial Super Intelligence

General Definition

- **A possible future level of AI that surpasses human intelligence in all aspects, including cognitive abilities, problem-solving, creativity, and emotional understanding**
- Already happen in some areas (e.g., protein folding, chess, go, etc.)
- May never happen in totality
- Many experts say it’s happening within a few years

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**

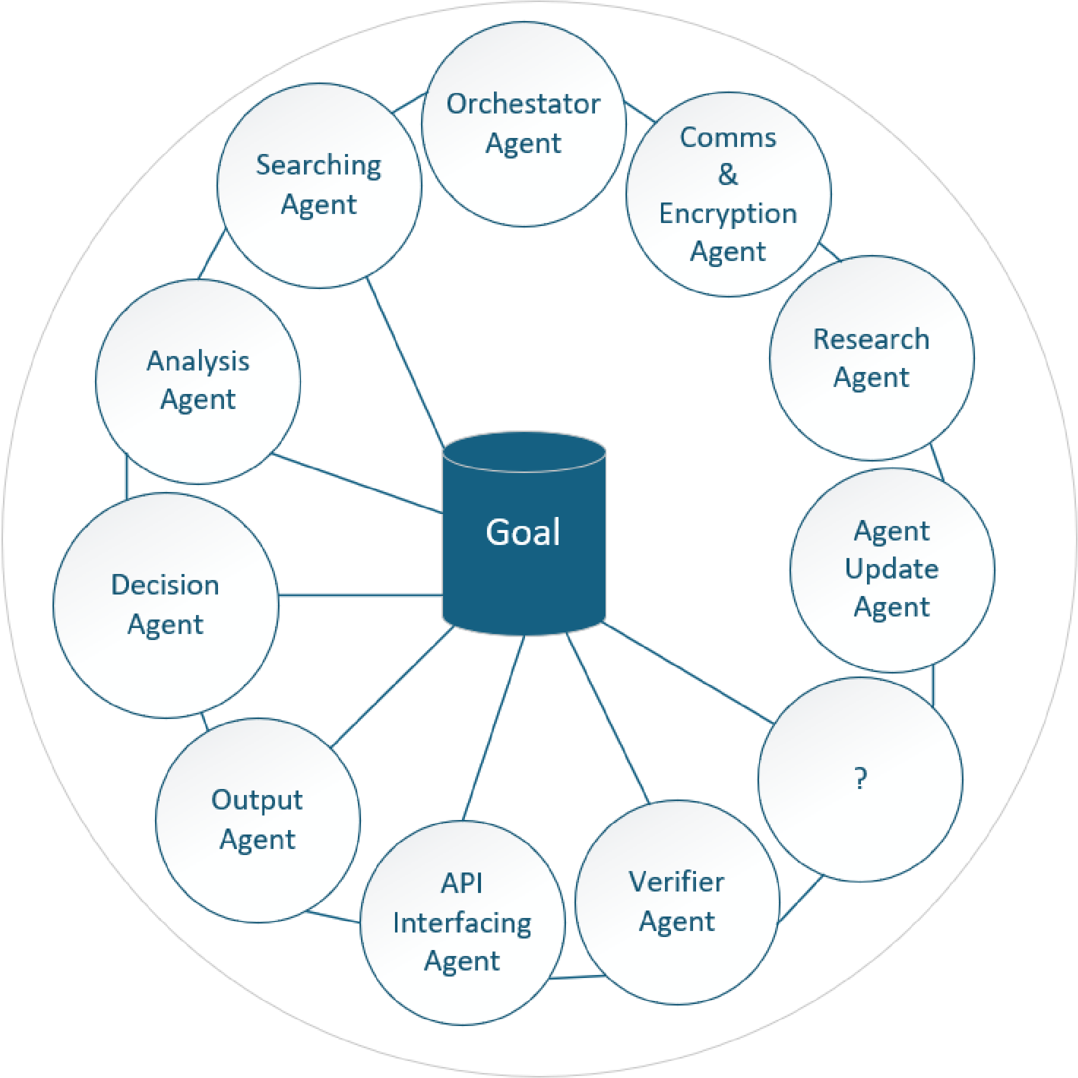
- Real-world allegory: Like building a house

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

What Is Agentic AI?

Agentic AI

Mock Agentic AI Template



What Is Agentic AI?

Automous 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!!**

What Is Agentic AI?

Automous Agentic AI

Requirements (for my definition):

- Involves AI-enabled LLM
- Able to change outputs based on inputs
- Cooperative agents
- Some level of autonomy
- Decides, creates, or does something

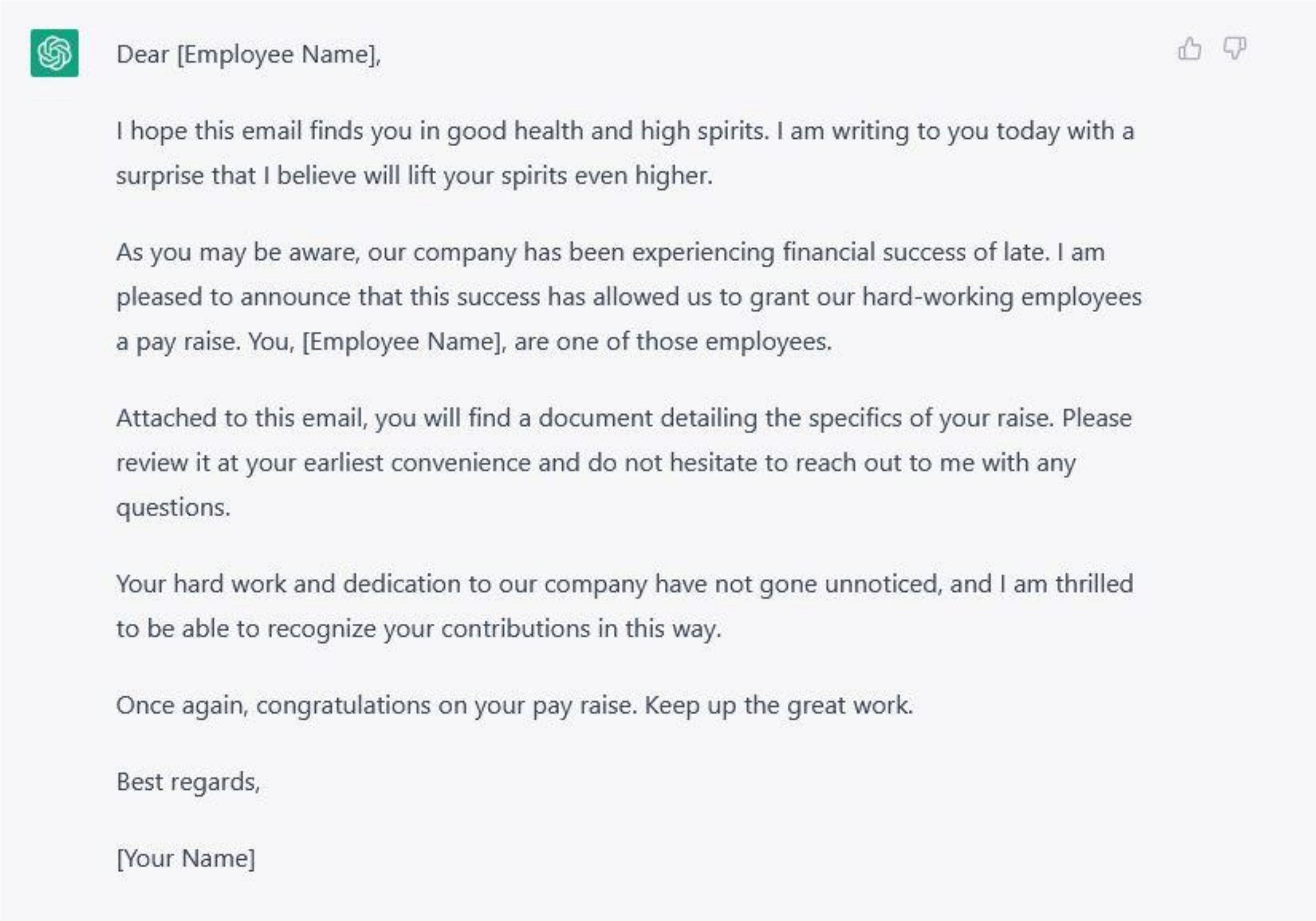
- Today's AI's shows things, but human still does the doing
- Future AI will create and do things
  - AI will do…what we can do…+robotics, biotech, nanotech, and 3D printing on steroids

Agenda

KnowBe4

- Traditional Hacking Attacks
- What Is AI and Agentic AI?
- Agentic AI Attacks
- 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

Old School

Fake Emails

Fake Images

Fake Audio

Fake Video of Fake People

Fake Video of Real People

Now & Improving

Fake Real-Time Video of Fake People

Fake Real-Time Video of Real People

Automated AI Deepfake Malware

Coming Soon

Autonomous Agentic AI Deepfake Malware

Malicious AI Threats

<u>Overview</u>

- Can be used to create very realistic phishing attacks
- We are already seeing very good AI “deepfakes” used to exploit people
  - “Deepfake” is very realistic, but fake picture, video, or audio recording of a person

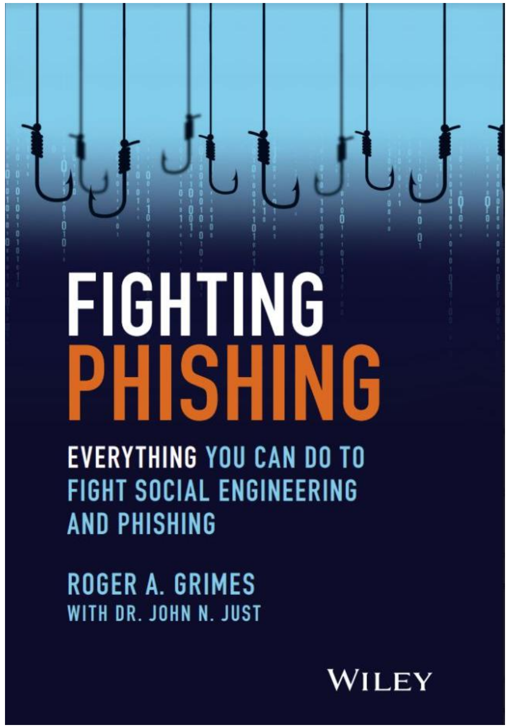
- AI used to craft very realistic-looking phishing emails
- AI used to respond to potential victims
- AI used to create “deepfakes” to fool victims
- Create new fake “synthetic” identities

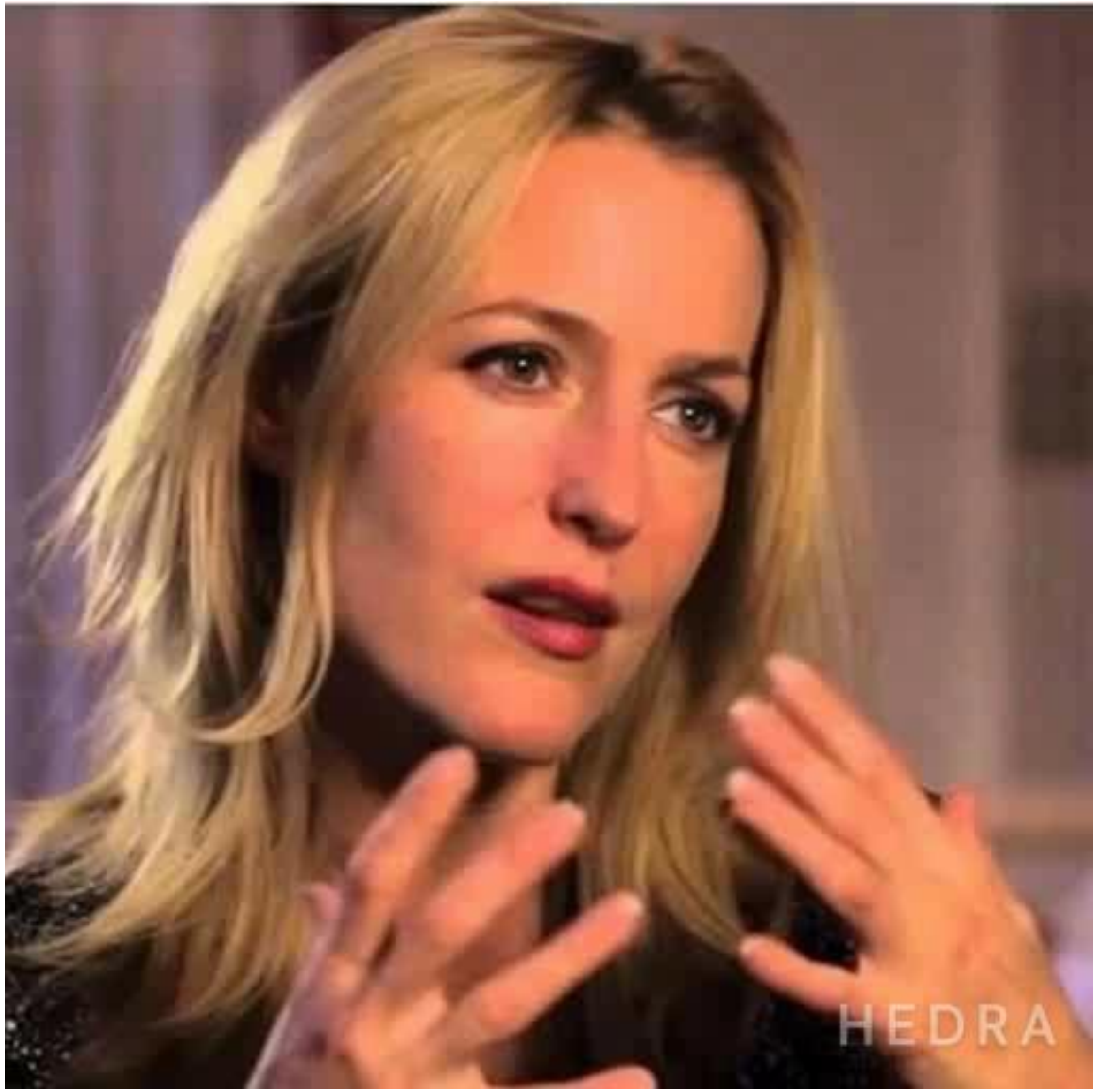
AI-Enabled Deepfakes

Summary

- Thousands of tools exist to help anyone create very realistic “deepfakes”
- Can make a real person say or look like they are do anything you want (digitally)
- Can make up a manipulated or brand new, never-existing-before person/image (known as a synthetic image)
- Pretty quickly and easily
- Tools are getting better every day

Apparently, actress
Gillian Anderson
loves my latest book

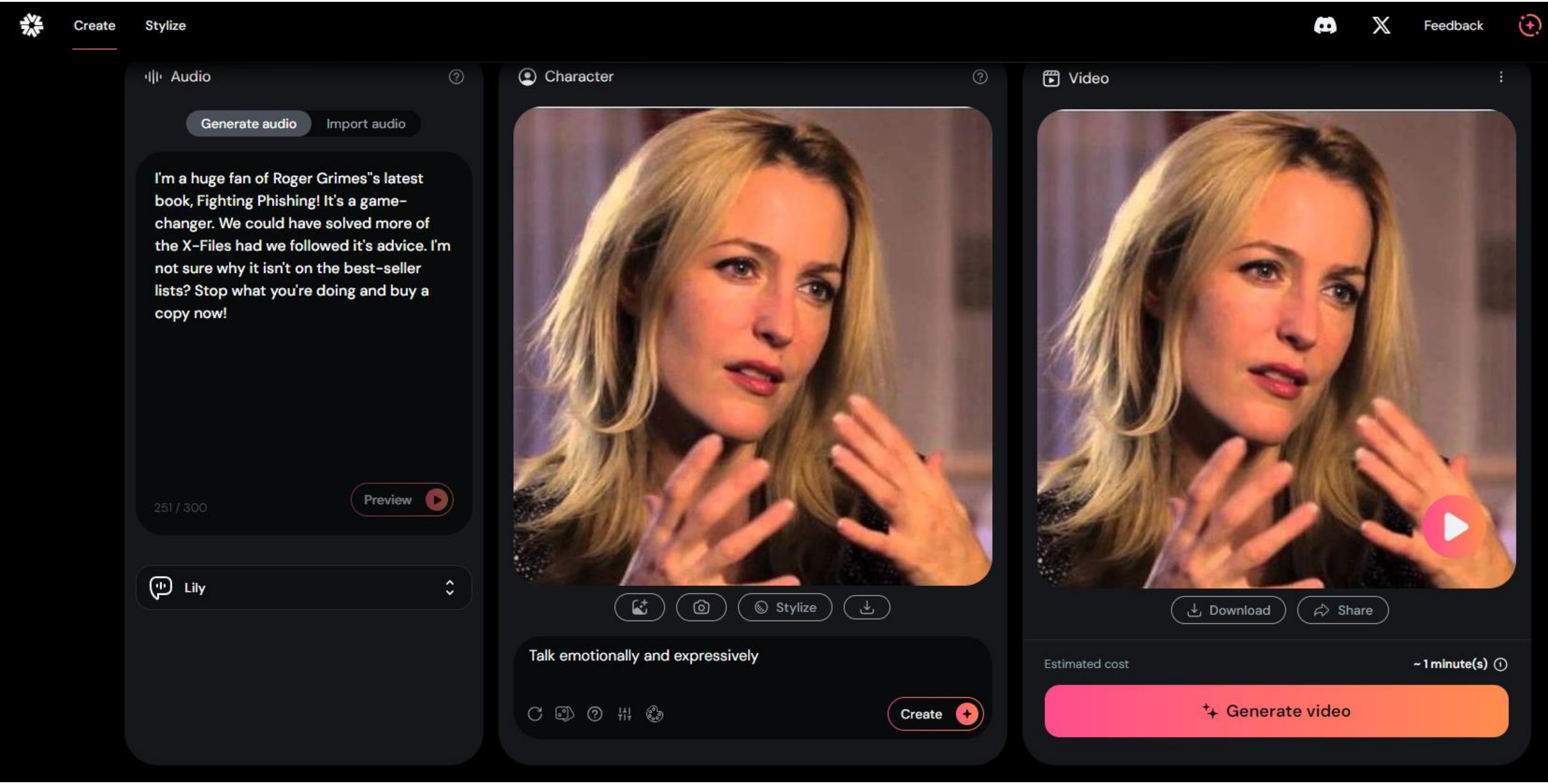




I did this in the first 30 seconds during my first time on an AI site, my first time ever trying to make a AI-created deepfake video

Imagine what anyone could do in 5 minutes?

https://www.hedra.com/



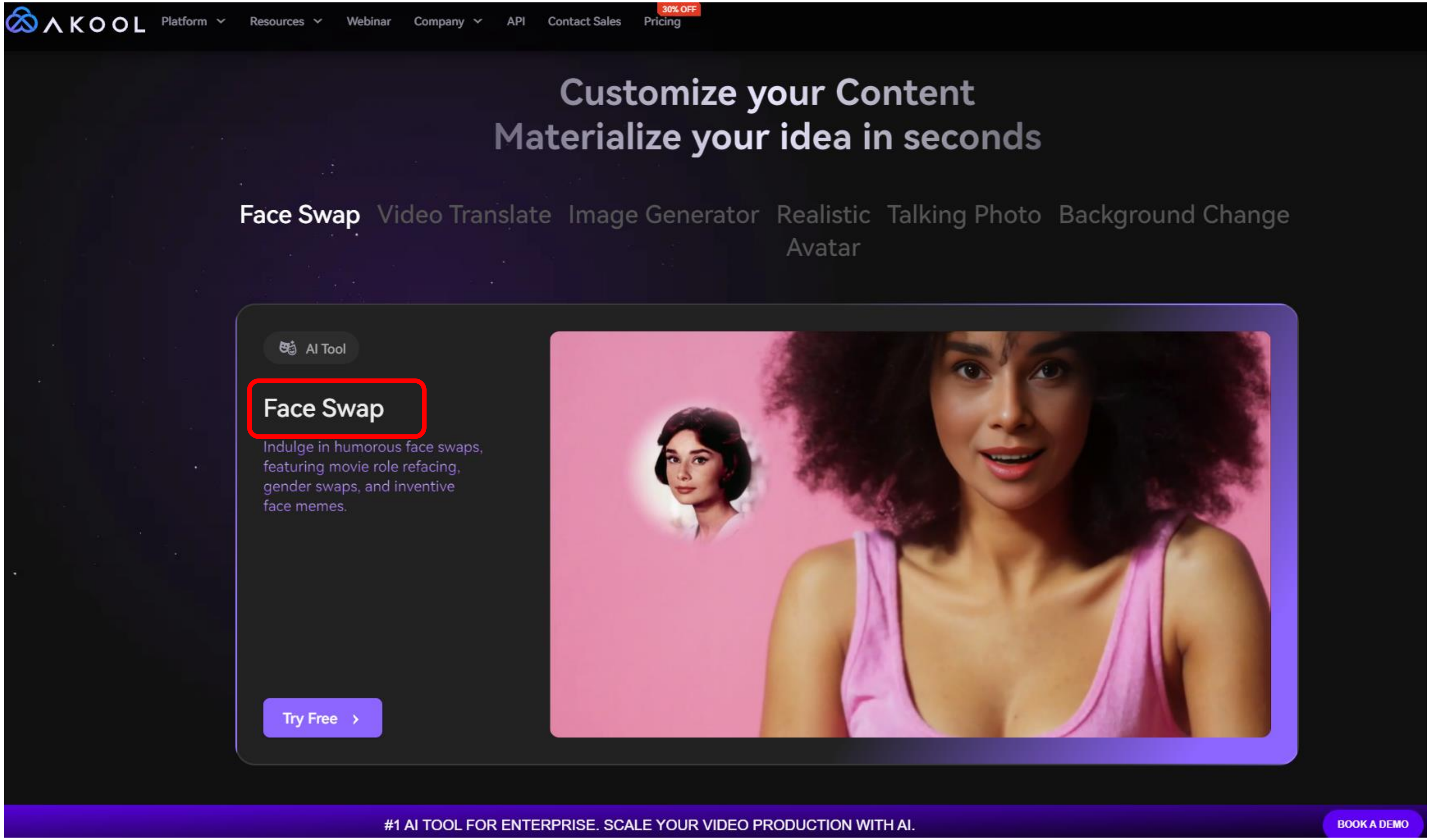
I'm a huge fan of Roger Grimes"s latest book, Fighting Phishing! It's a game-changer. We could have solved more of the X-Files had we followed it's advice. I'm not sure why it isn't on the best-seller lists? Stop what you're doing and buy a copy now!

251 / 300

Talk emotionally and expressively

Estimated cost ~1 minute(s)

Deepfakes



Customize your Content
Materialize your idea in seconds

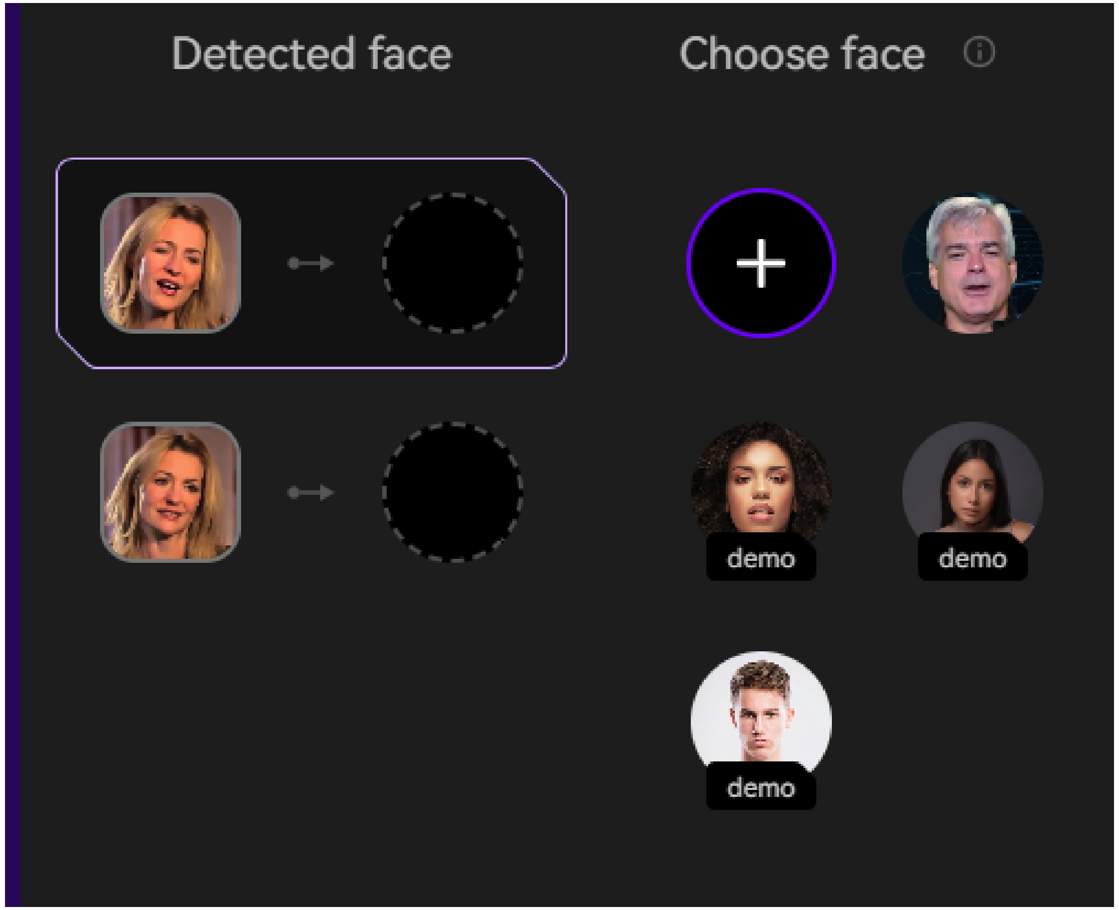
Face Swap

Indulge in humorous face swaps, featuring movie role refacing, gender swaps, and inventive face memes.

#1 AI TOOL FOR ENTERPRISE. SCALE YOUR VIDEO PRODUCTION WITH AI.

www.akool.com

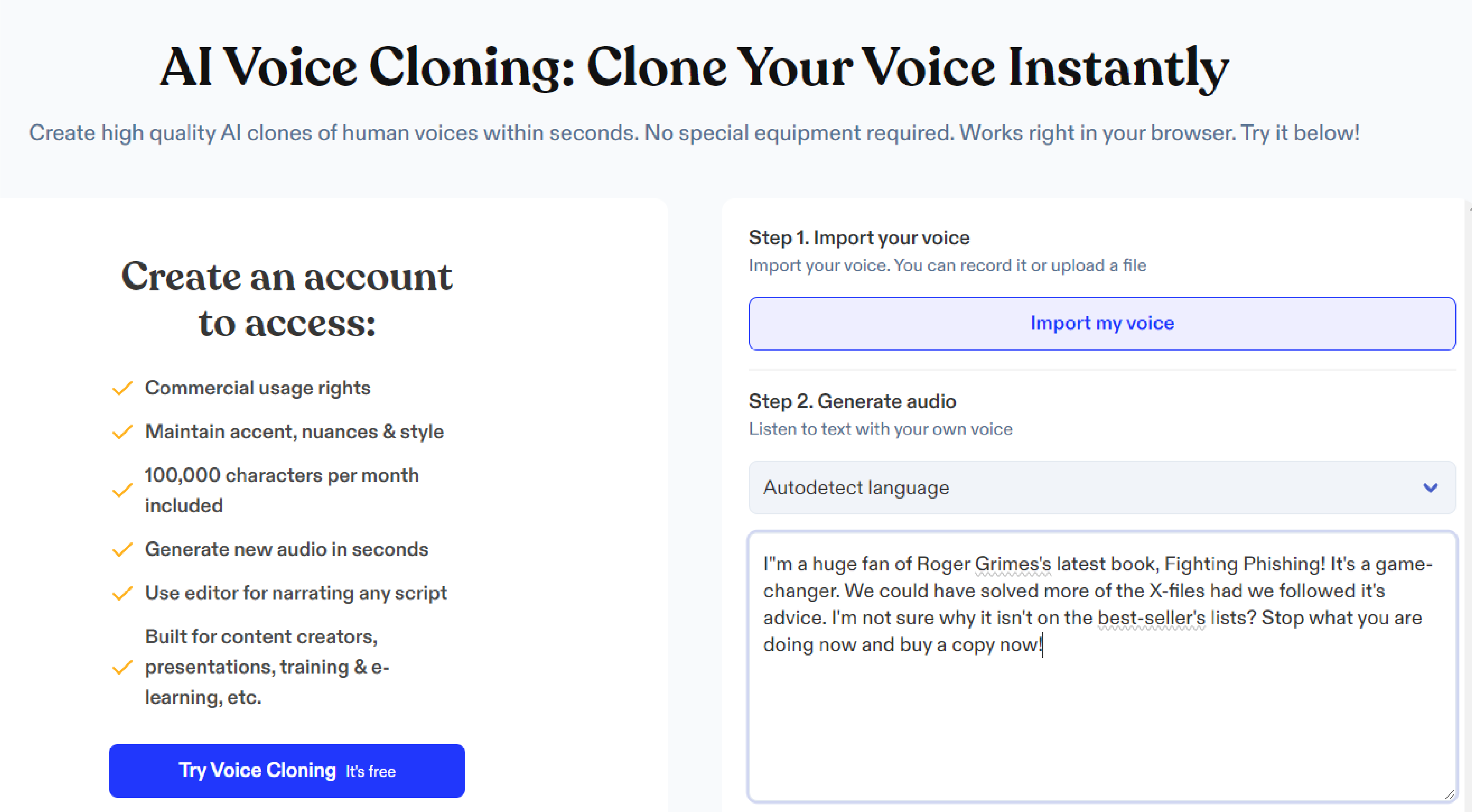
Deepfakes



Swapped Faces

www.akool.com

Deepfakes



AI Voice Cloning: Clone Your Voice Instantly

Create high quality AI clones of human voices within seconds. No special equipment required. Works right in your browser. Try it below!

Create an account to access:

- ✓ Commercial usage rights
- ✓ Maintain accent, nuances & style
- ✓ 100,000 characters per month included
- ✓ Generate new audio in seconds
- ✓ Use editor for narrating any script
- ✓ Built for content creators, presentations, training & e-learning, etc.

Try Voice Cloning It's free

Step 1. Import your voice

Import your voice. You can record it or upload a file

Import my voice

Step 2. Generate audio

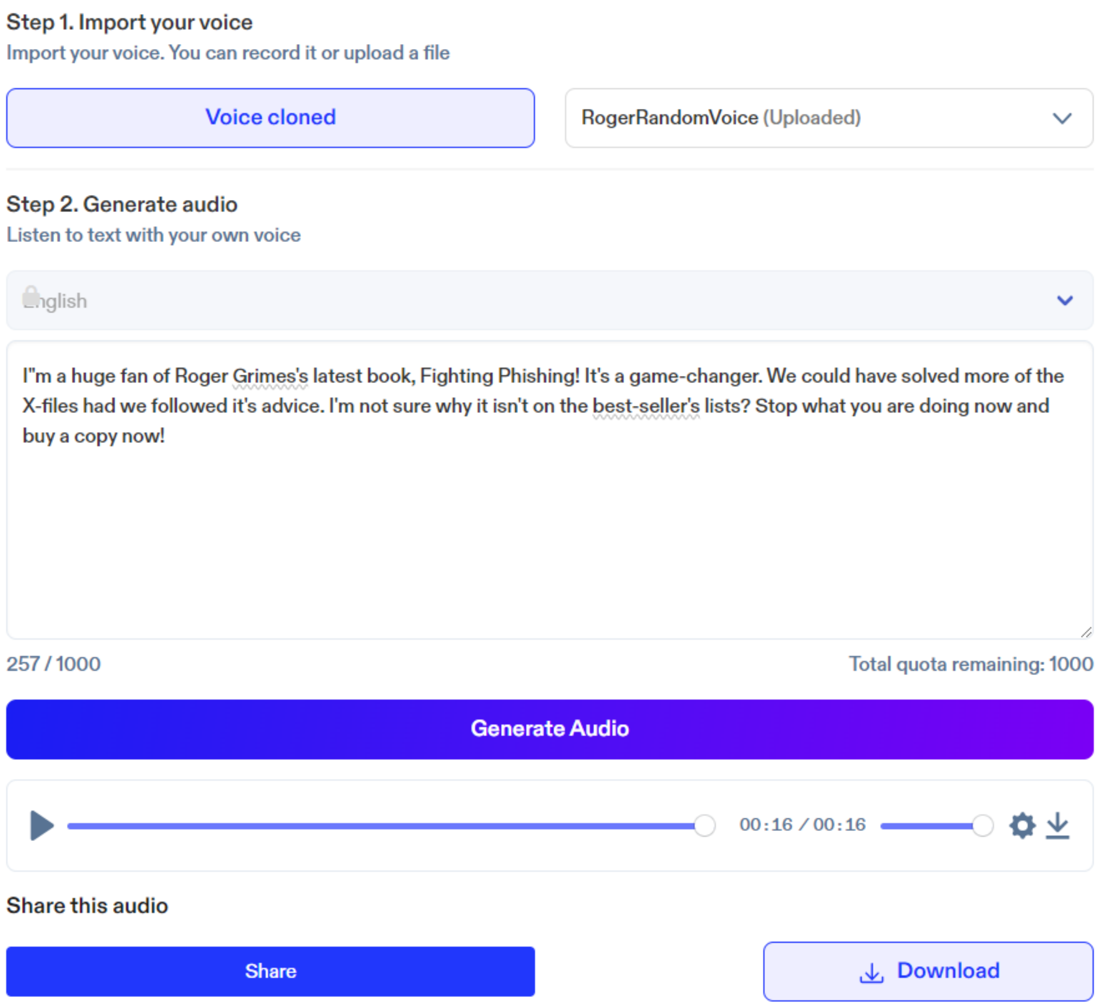
Listen to text with your own voice

Autodetect language

I"m a huge fan of Roger Grimes's latest book, Fighting Phishing! It's a game-changer. We could have solved more of the X-files had we followed it's advice. I'm not sure why it isn't on the best-seller's lists? Stop what you are doing now and buy a copy now!

https://myvoice.speechify.com/

Deepfakes



Step 1. Import your voice

Import your voice. You can record it or upload a file

RogerRandomVoice (Uploaded)

Step 2. Generate audio

Listen to text with your own voice

English

I"m a huge fan of Roger Grimes's latest book, Fighting Phishing! It's a game-changer. We could have solved more of the X-files had we followed it's advice. I'm not sure why it isn't on the best-seller's lists? Stop what you are doing now and buy a copy now!

257 / 1000

Total quota remaining: 1000

00:16 / 00:16

Share this audio

Uploaded an audio clip of my voice saying something else

Rough Draft

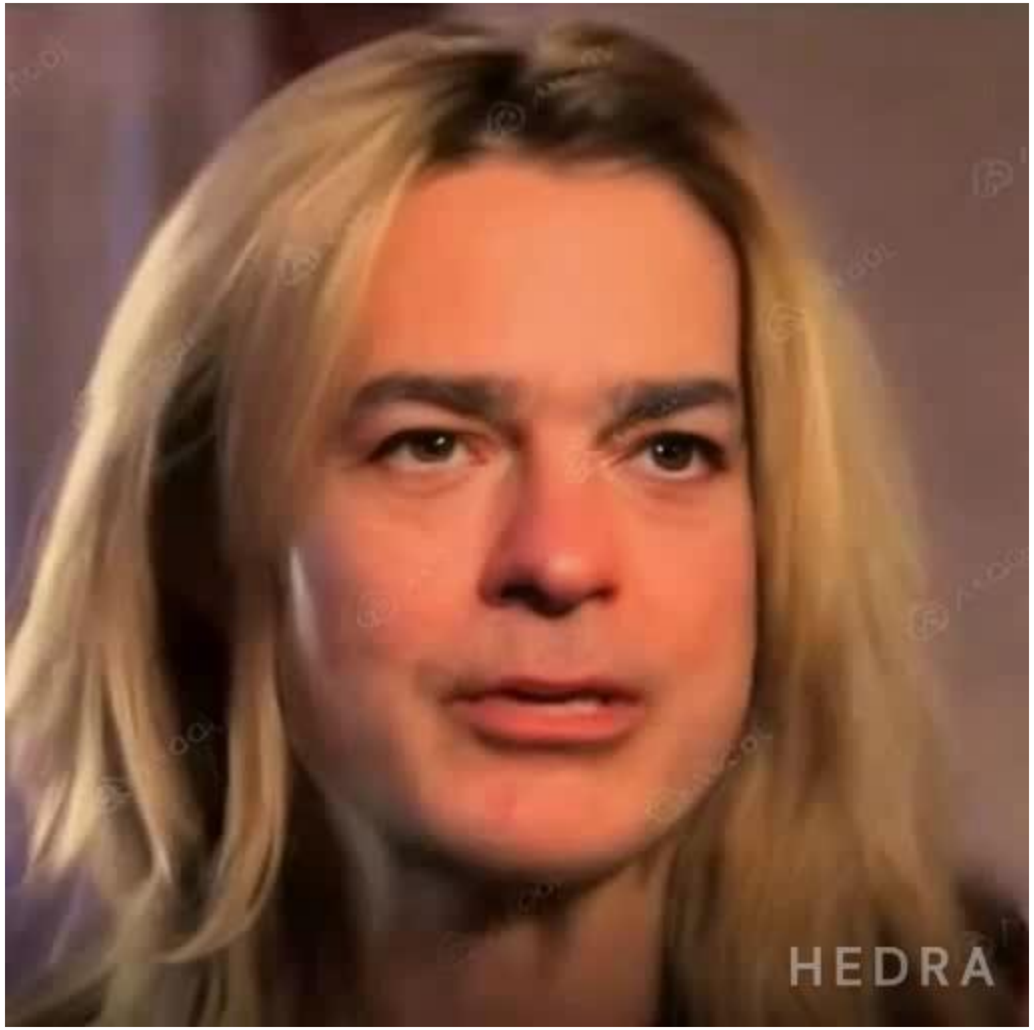
Final Version

https://myvoice.speechify.com/

Deepfakes

Face Swap
&
AI-Generated Voice

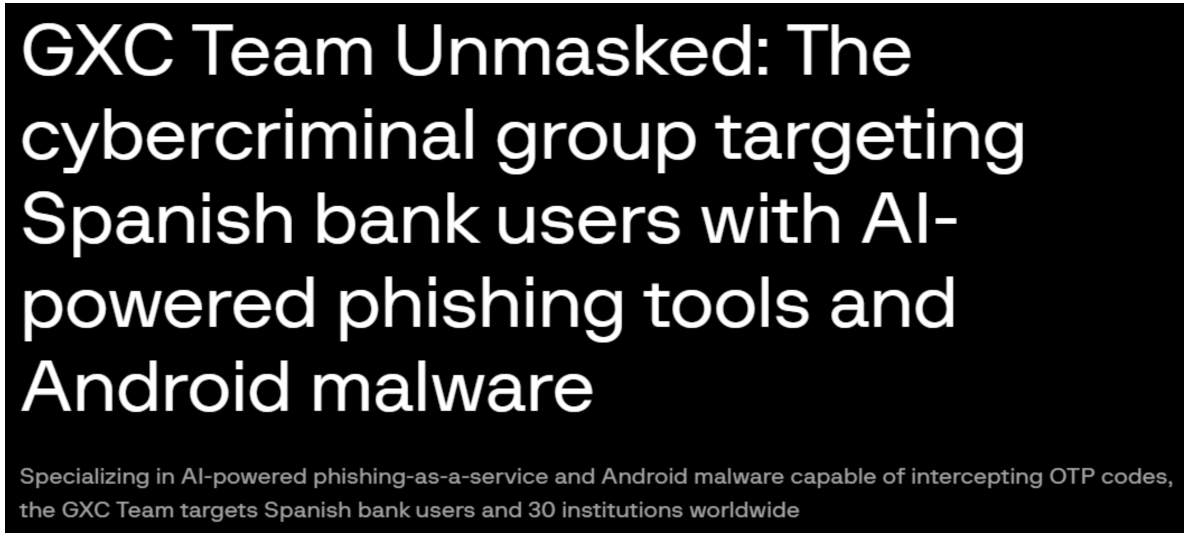
Uploaded Face
Swapped Picture
Uploaded Fake Audio
Took 30 seconds



https://www.hedra.com/

AI Used To Fake Bank Calls

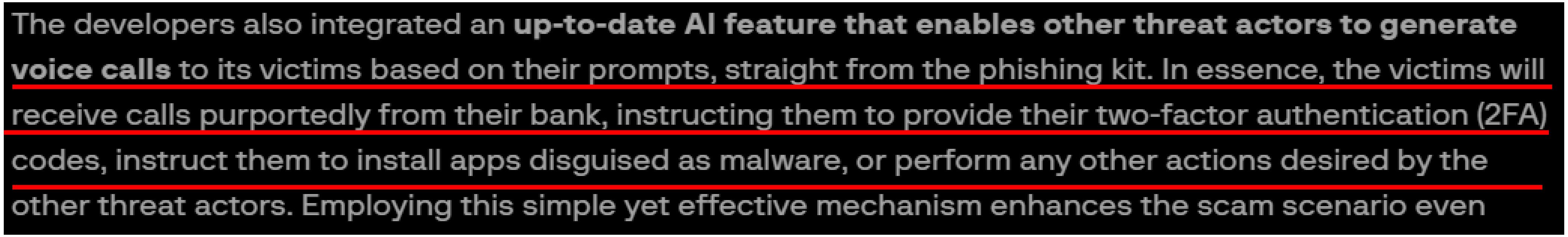
Fake Bank Calls



GXC Team Unmasked: The cybercriminal group targeting Spanish bank users with AI-powered phishing tools and Android malware

Specializing in AI-powered phishing-as-a-service and Android malware capable of intercepting OTP codes, the GXC Team targets Spanish bank users and 30 institutions worldwide

AI-powered voice caller feature in the phishing kit

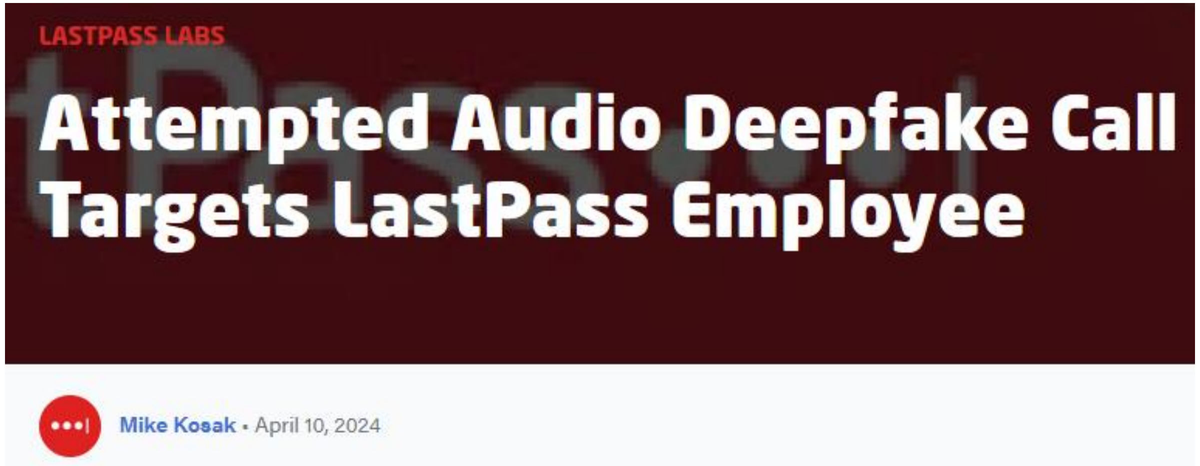


The developers also integrated an **up-to-date AI feature that enables other threat actors to generate voice calls** to its victims based on their prompts, straight from the phishing kit. In essence, the victims will receive calls purportedly from their bank, instructing them to provide their two-factor authentication (2FA) codes, instruct them to install apps disguised as malware, or perform any other actions desired by the other threat actors. Employing this simple yet effective mechanism enhances the scam scenario even

https://www.group-ib.com/blog/gxc-team-unmasked/

AI-Created Deepfakes Used In Attempt Theft

Fake Calls

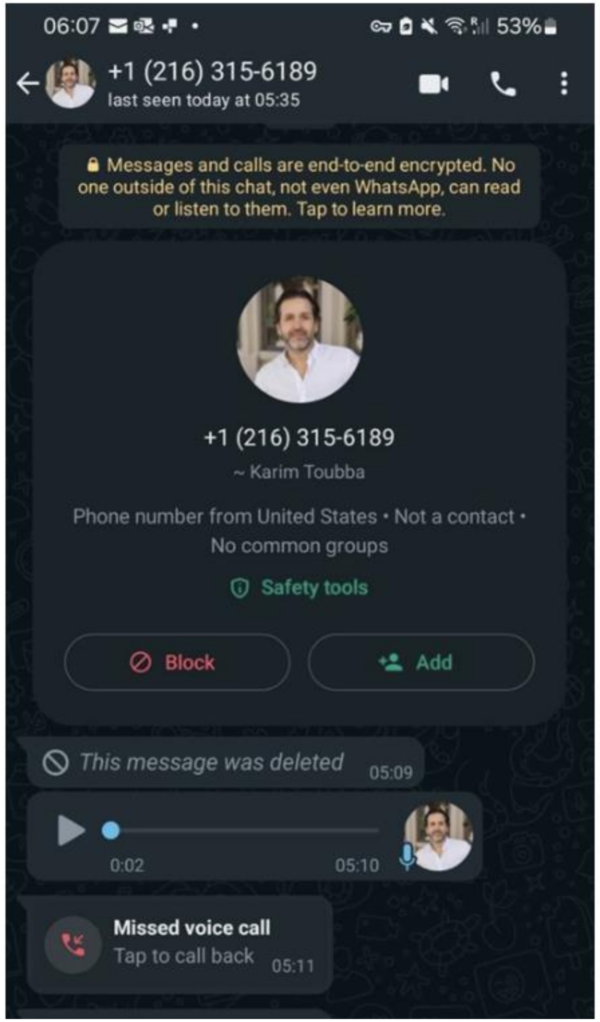


LASTPASS LABS

Attempted Audio Deepfake Call Targets LastPass Employee

Mike Kosak • April 10, 2024

Mike Kosak, Senior Principal Intelligence Analyst at LastPass, explained in a blog post, "In our case, an employee received a series of calls, texts, and at least one voicemail featuring an audio deepfake from a threat actor impersonating our CEO via WhatsApp.



06:07 53%

+1 (216) 315-6189
last seen today at 05:35

🔒 Messages and calls are end-to-end encrypted. No one outside of this chat, not even WhatsApp, can read or listen to them. Tap to learn more.

+1 (216) 315-6189
~ Karim Toubba

Phone number from United States • Not a contact • No common groups

Safety tools

This message was deleted 05:09

0:02 05:10

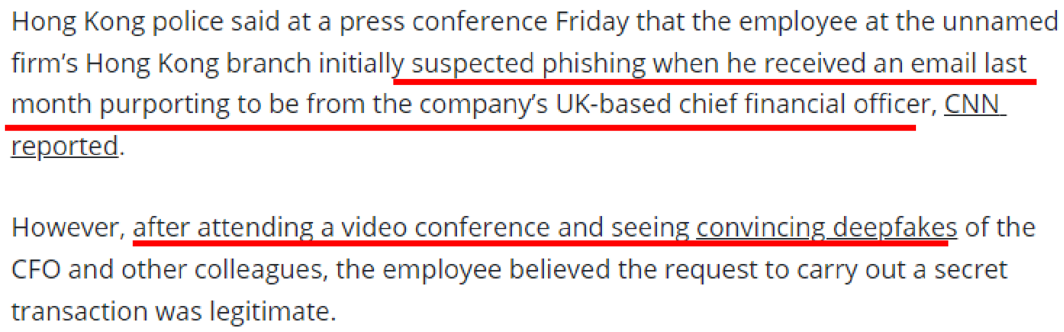
Missed voice call
Tap to call back 05:11

https://blog.lastpass.com/posts/2024/04/attempted-audio-deepfake-call-targets-lastpass-employee

Examples

AI/Deepfakes

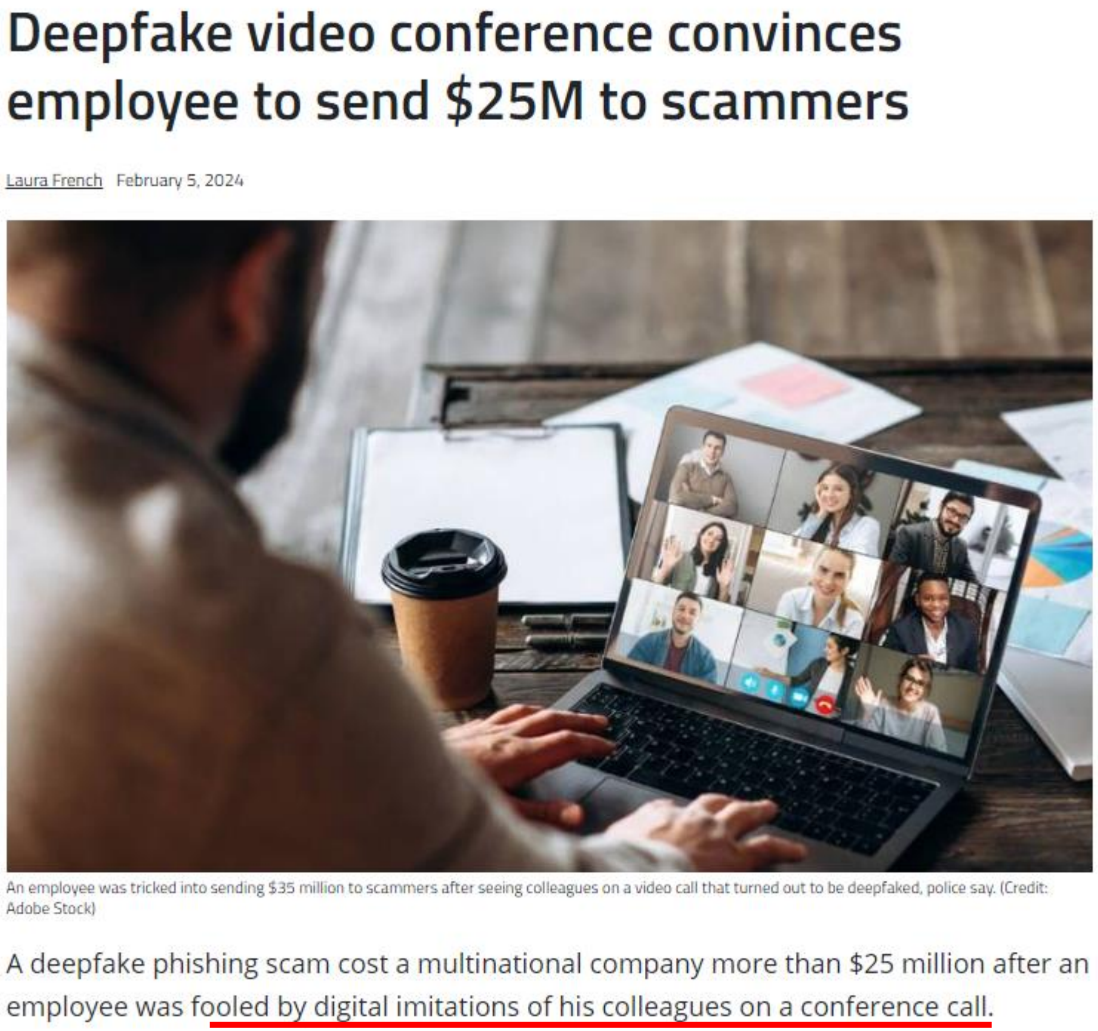
Just making long-game phishing easier to do



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

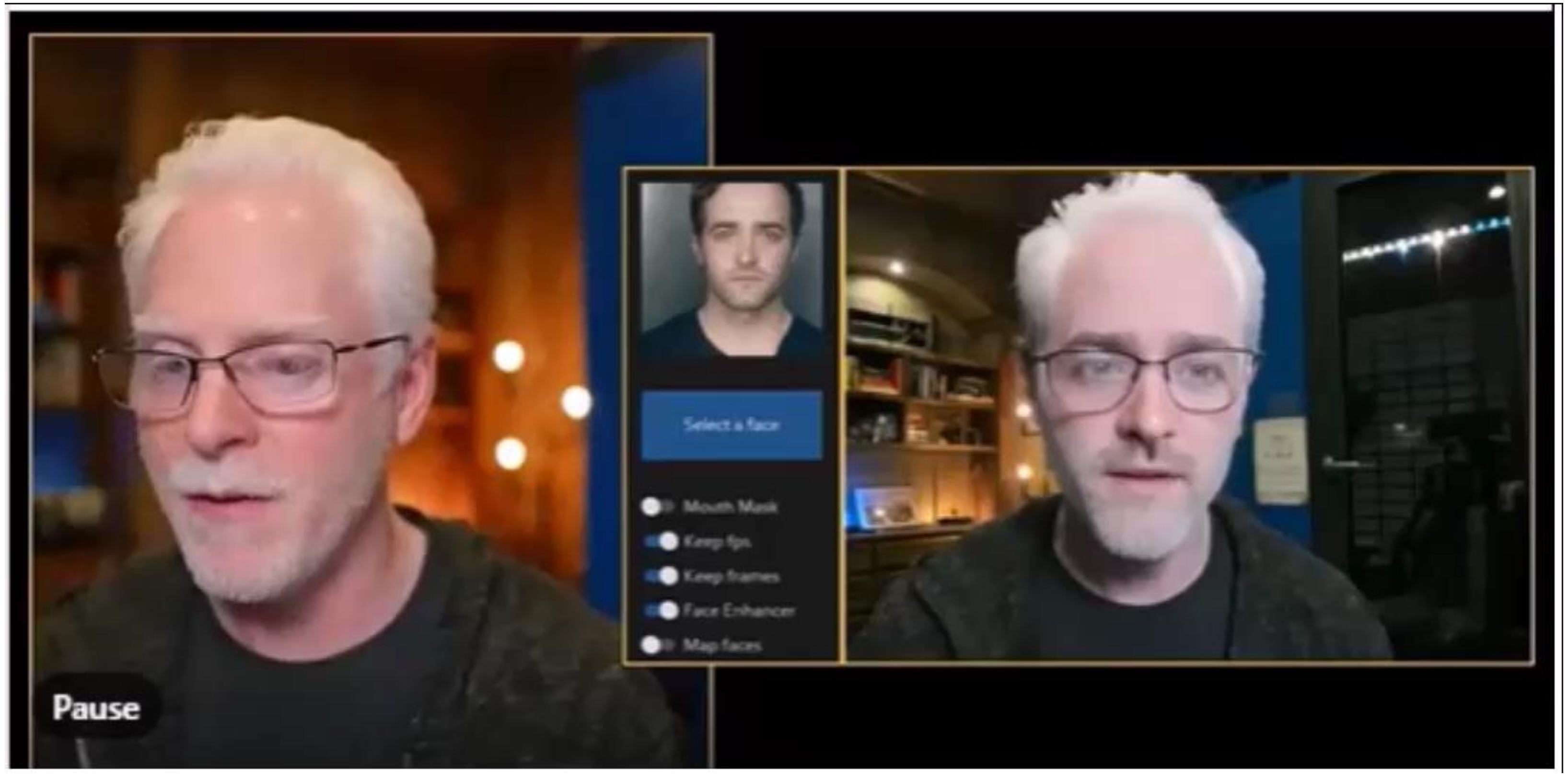
An employee was tricked into sending $35 million to scammers after seeing colleagues on a video call that turned out to be deepfaked, police say. (Credit: Adobe Stock)

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



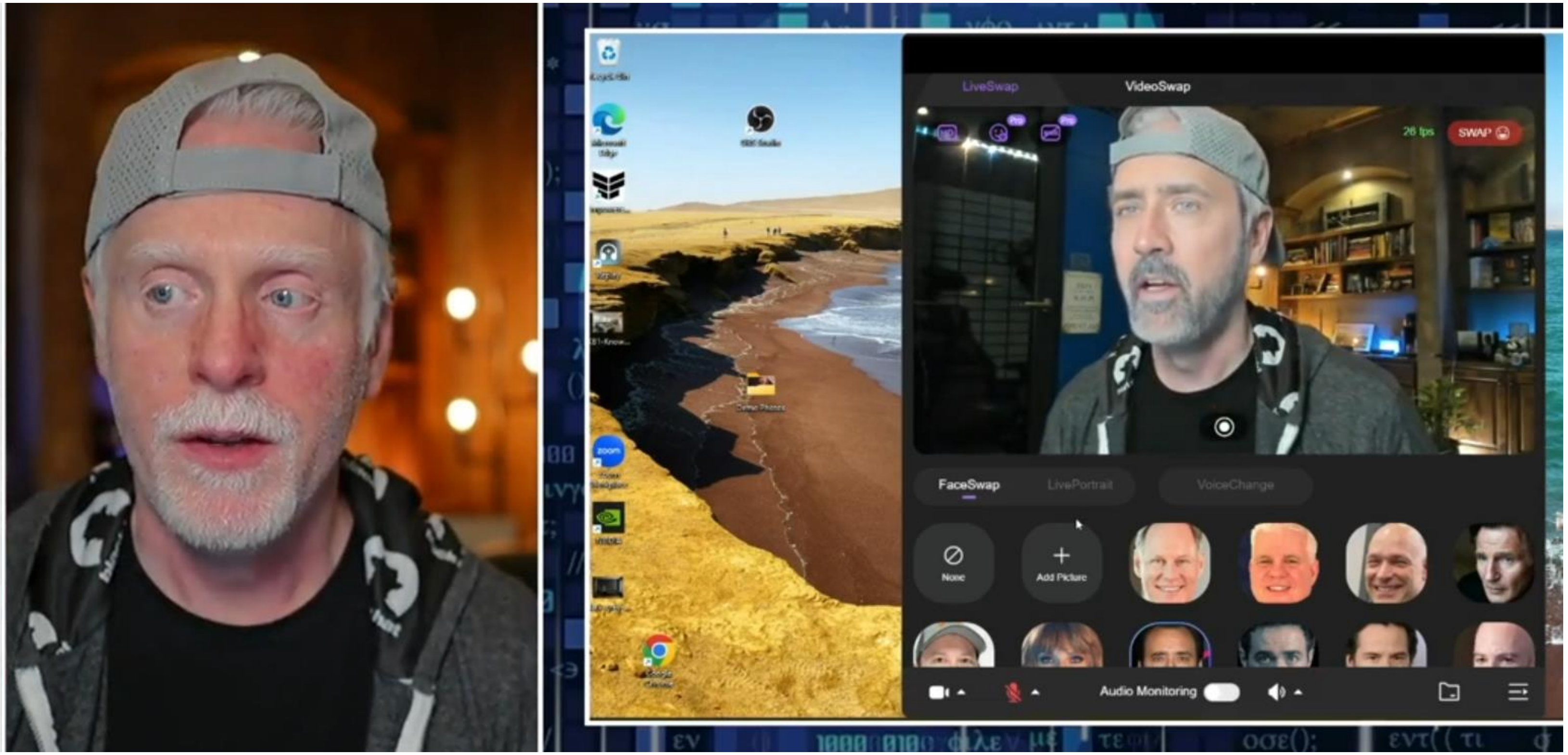
- ○ Mouth Mask
- ● Keep fps
- ● Keep frames
- ● Face Enhancer
- ○ Map faces

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

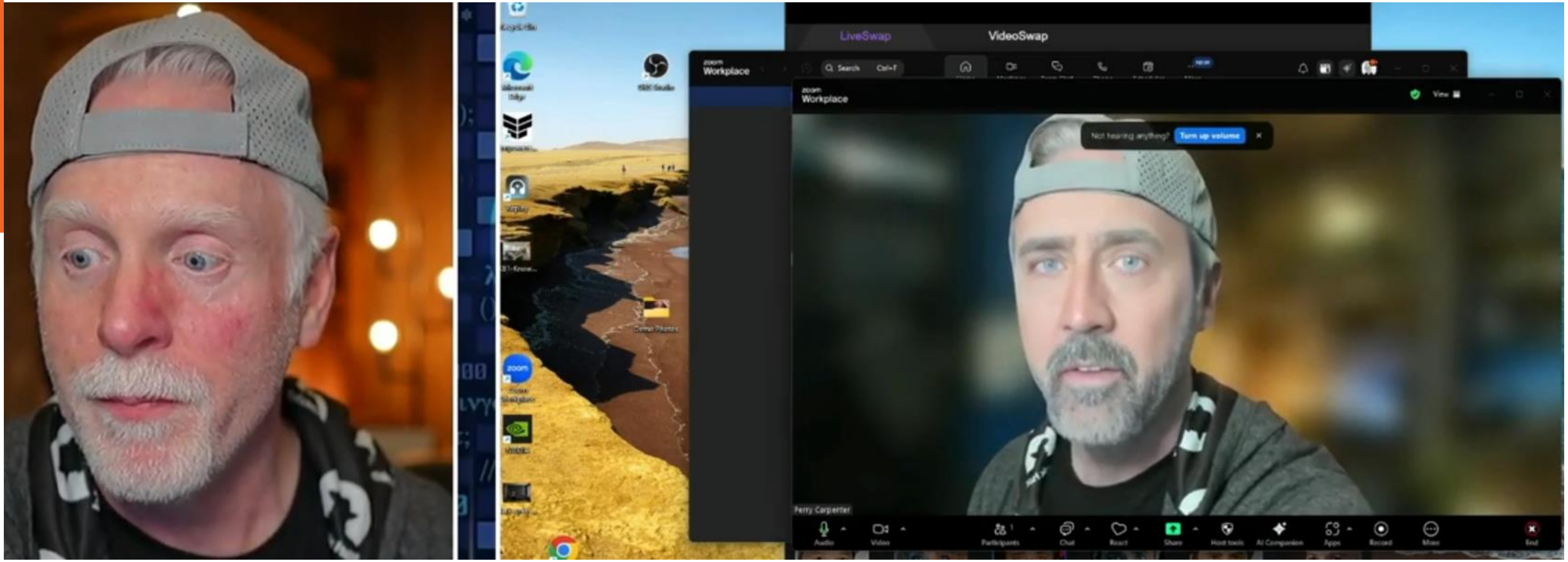


https://www.linkedin.com/posts/perrycarpenter_cybersecurity-securityawareness-ai-activity-7318408334393384960-_vQD

Real-Time AI-Enabled Deepfakes

Fake Real-Time Video

Deepfaking Zoom Calls

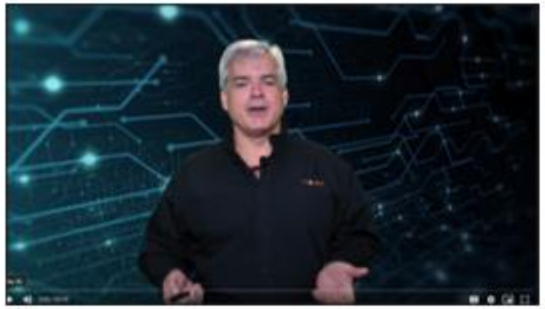


https://www.linkedin.com/posts/perrycarpenter_cybersecurity-securityawareness-ai-activity-7318408334393384960-_vQD

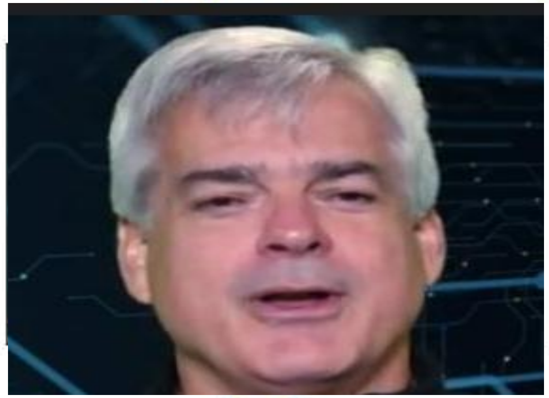
It’s A Hacker Industry Now

Reused Stolen Biometrics

Step 1:
Steal Biometric Attribute (e.g., face, fingerprint, voice, video, etc.) from victim or video



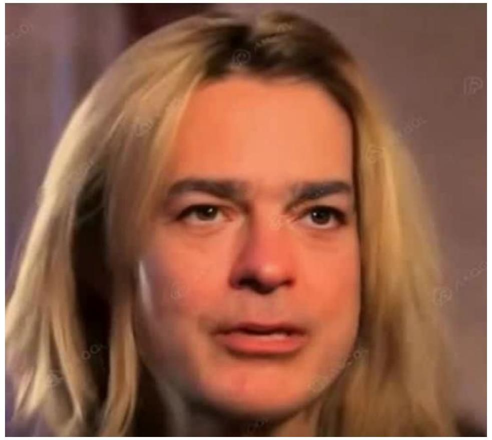
Step 2:
Create Deepfake



Step 3:
Use Device Emulator to emulate victim’s device (e.g., OS, browser, location, etc.)

Step 4:
Use Virtual Camera
(“camera software” that allows deepfake injection)

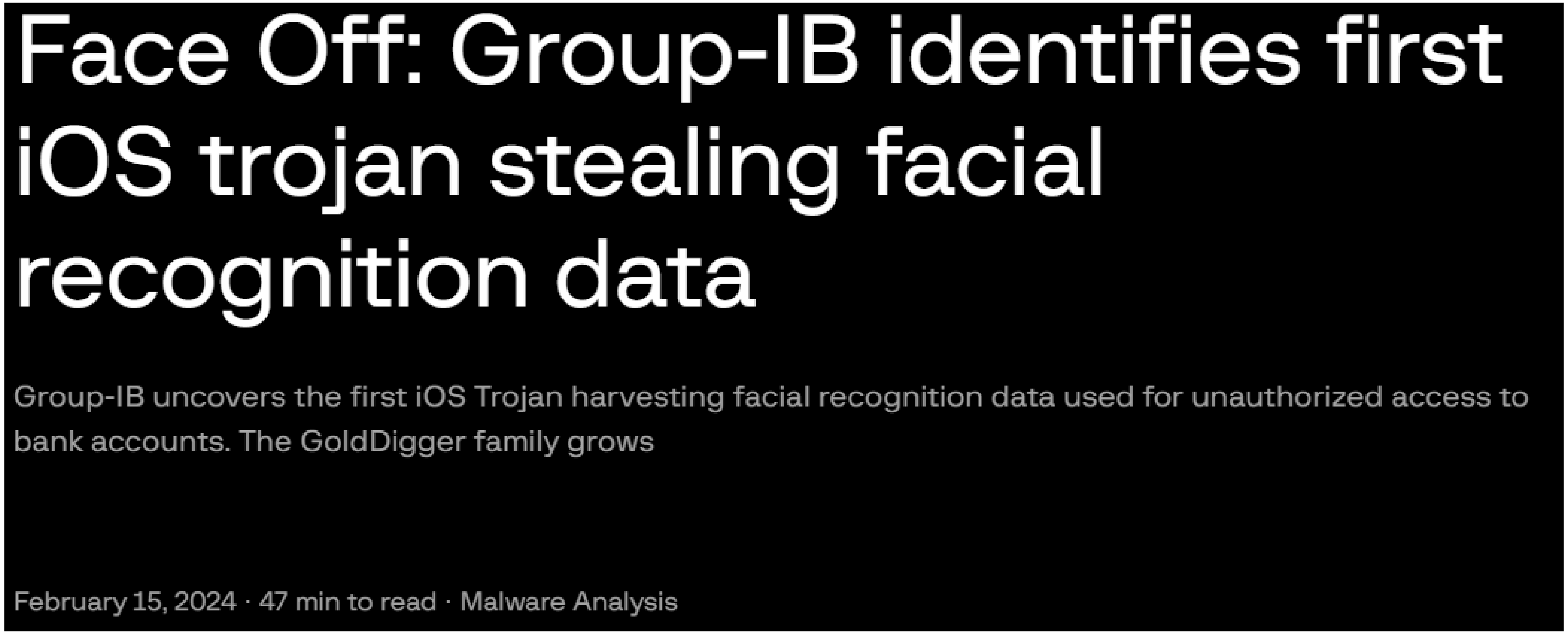
Step 5
Perform Deepfake Attack



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

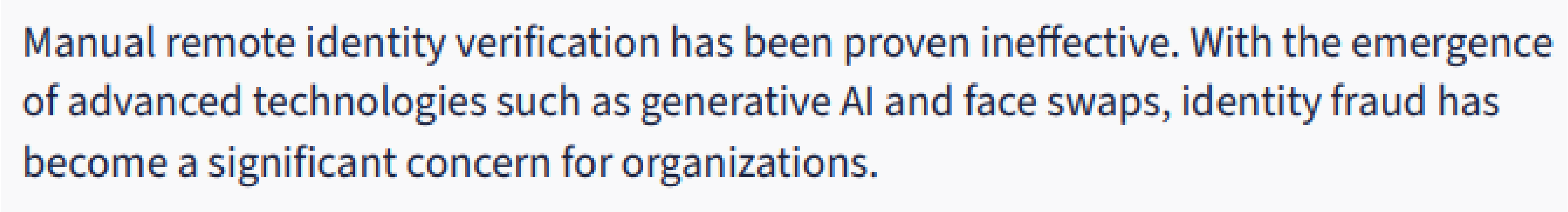
February 15, 2024 · 47 min to read · Malware Analysis

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

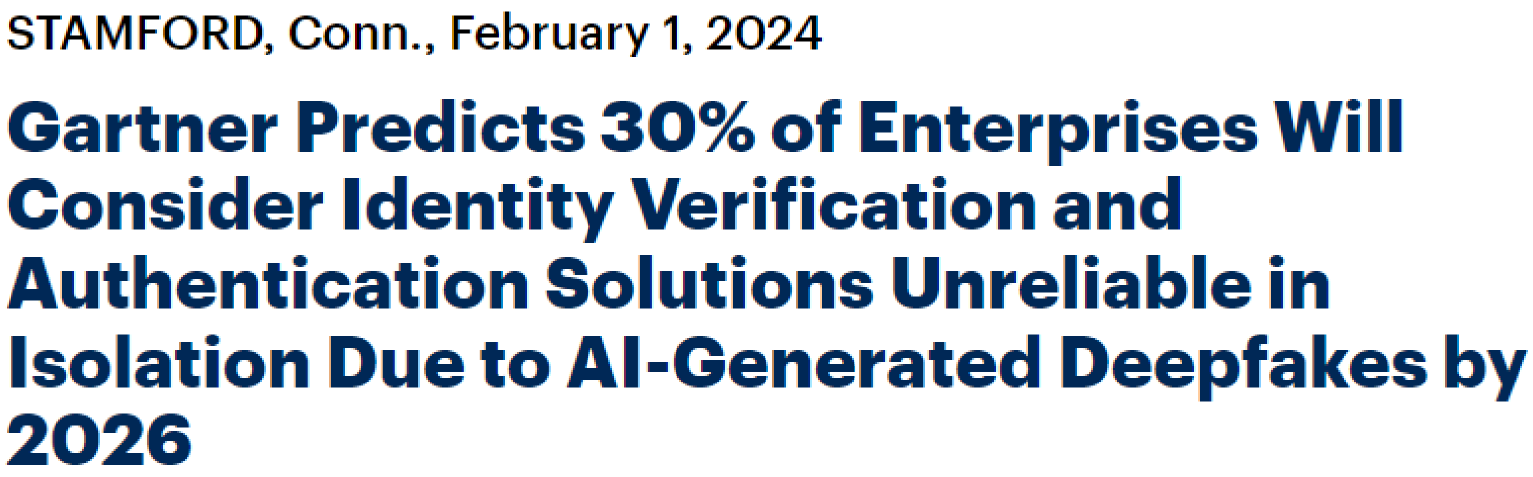
https://www.group-ib.com/blog/goldfactory-ios-trojan/

Can You Trust Biometrics?

Reused
Stolen
Biometrics



Manual remote identity verification has been proven ineffective. With the emergence of advanced technologies such as generative AI and face swaps, identity fraud has become a significant concern for organizations.



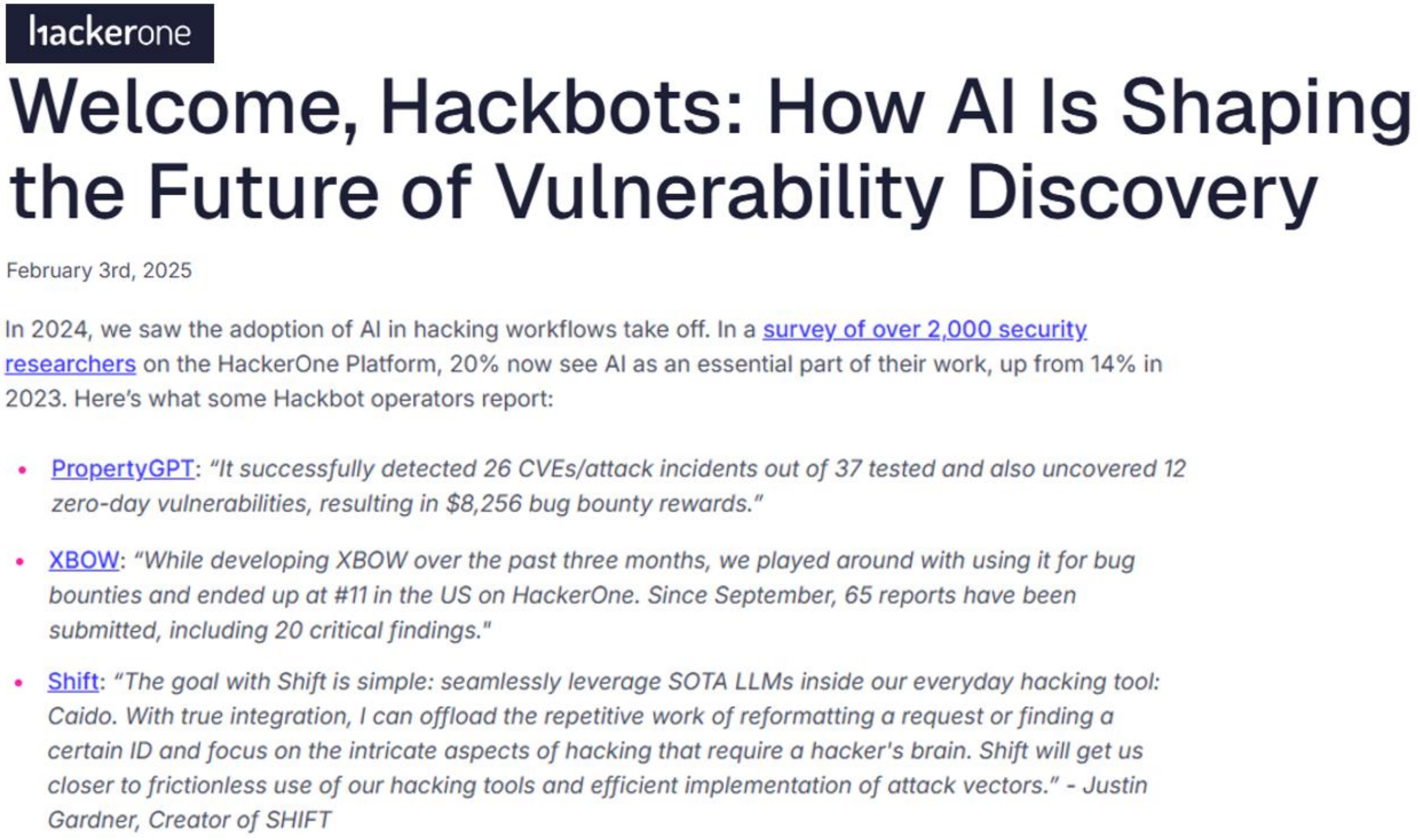
STAMFORD, Conn., February 1, 2024

Gartner Predicts 30% of Enterprises Will Consider Identity Verification and Authentication Solutions Unreliable in Isolation Due to AI-Generated Deepfakes by 2026

https://www.gartner.com/en/newsroom/press-releases/2024-02-01-gartner-predicts-30-percent-of-enterprises-will-consider-identity-verification-and-authentication-solutions-unreliable-in-isolation-due-to-deepfakes-by-2026

Hackbots

AI-enabled bots that hack things



hackerone

Welcome, Hackbots: How AI Is Shaping the Future of Vulnerability Discovery

February 3rd, 2025

In 2024, we saw the adoption of AI in hacking workflows take off. In a survey of over 2,000 security researchers on the HackerOne Platform, 20% now see AI as an essential part of their work, up from 14% in 2023. Here's what some Hackbot operators report:

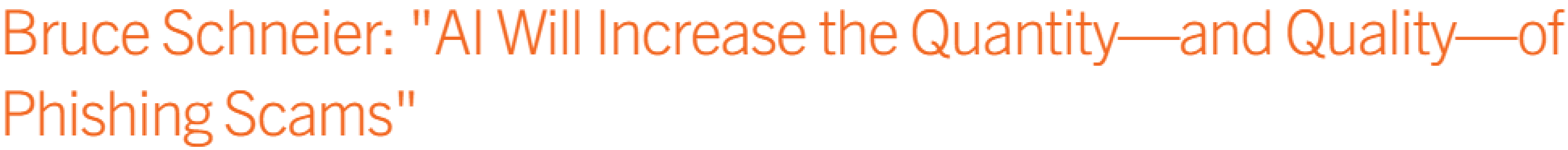
- PropertyGPT: *"It successfully detected 26 CVEs/attack incidents out of 37 tested and also uncovered 12 zero-day vulnerabilities, resulting in $8,256 bug bounty rewards."*
- XBOW: *"While developing XBOW over the past three months, we played around with using it for bug bounties and ended up at #11 in the US on HackerOne. Since September, 65 reports have been submitted, including 20 critical findings."*
- Shift: *"The goal with Shift is simple: seamlessly leverage SOTA LLMs inside our everyday hacking tool: Caido. With true integration, I can offload the repetitive work of reformatting a request or finding a certain ID and focus on the intricate aspects of hacking that require a hacker's brain. Shift will get us closer to frictionless use of our hacking tools and efficient implementation of attack vectors." - Justin Gardner, Creator of SHIFT*

https://www.hackerone.com/blog/welcome-hackbots-how-ai-shaping-future-vulnerability-discovery

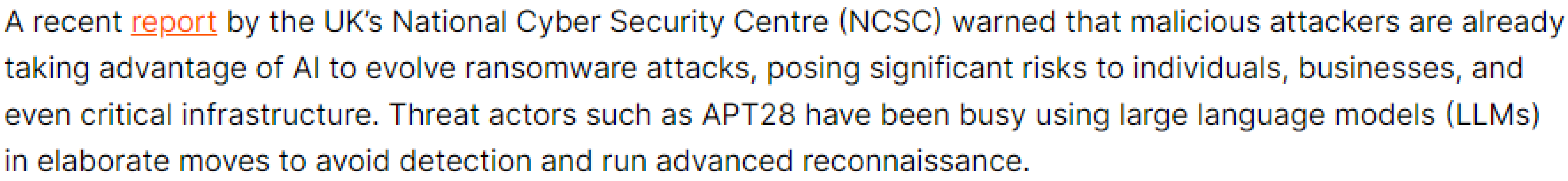
AI Attacks

Is AI Helping Attacks?

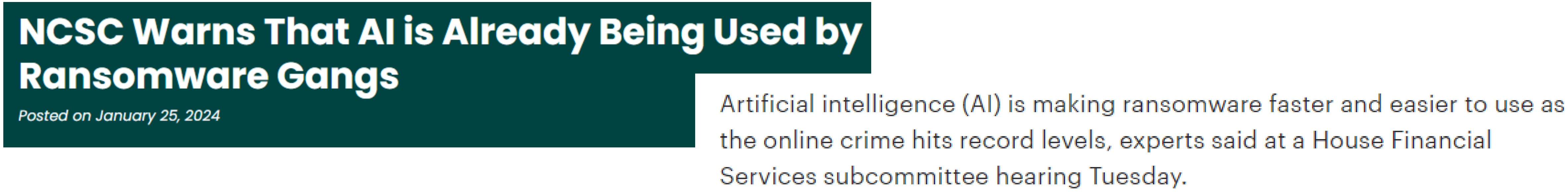
Yes, but it’s minor right now, but will absolutely be more over time



Bruce Schneier: "AI Will Increase the Quantity—and Quality—of Phishing Scams"



A recent report by the UK's National Cyber Security Centre (NCSC) warned that malicious attackers are already taking advantage of AI to evolve ransomware attacks, posing significant risks to individuals, businesses, and even critical infrastructure. Threat actors such as APT28 have been busy using large language models (LLMs) in elaborate moves to avoid detection and run advanced reconnaissance.



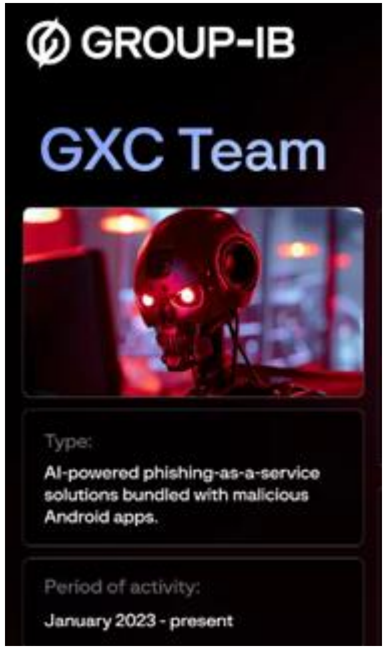
NCSC Warns That AI is Already Being Used by Ransomware Gangs

Posted on January 25, 2024

Artificial intelligence (AI) is making ransomware faster and easier to use as the online crime hits record levels, experts said at a House Financial Services subcommittee hearing Tuesday.

AI-Enabled Malware Is Getting Better

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



GROUP-IB

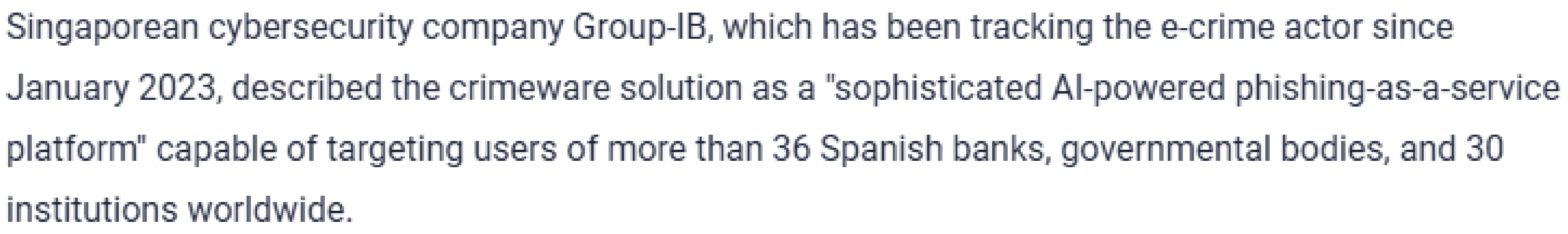
GXC Team

Type:

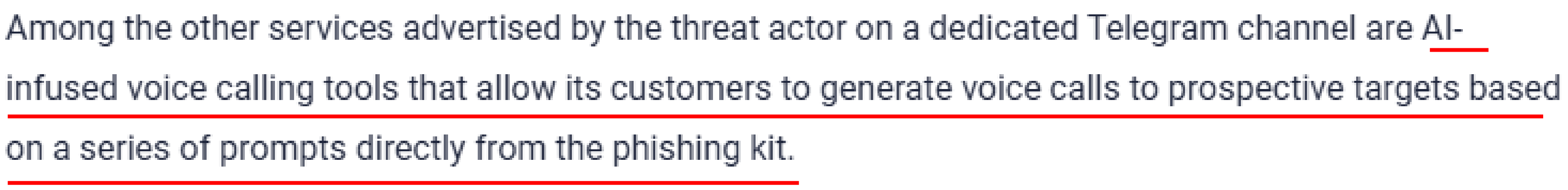
AI-powered phishing-as-a-service solutions bundled with malicious Android apps.

Period of activity:

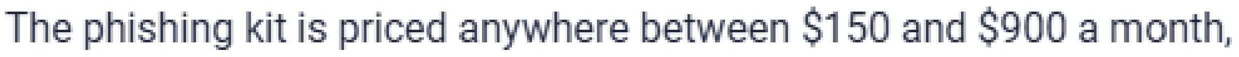
January 2023 - present



Singaporean cybersecurity company Group-IB, which has been tracking the e-crime actor since January 2023, described the crimeware solution as a "sophisticated AI-powered phishing-as-a-service platform" capable of targeting users of more than 36 Spanish banks, governmental bodies, and 30 institutions worldwide.



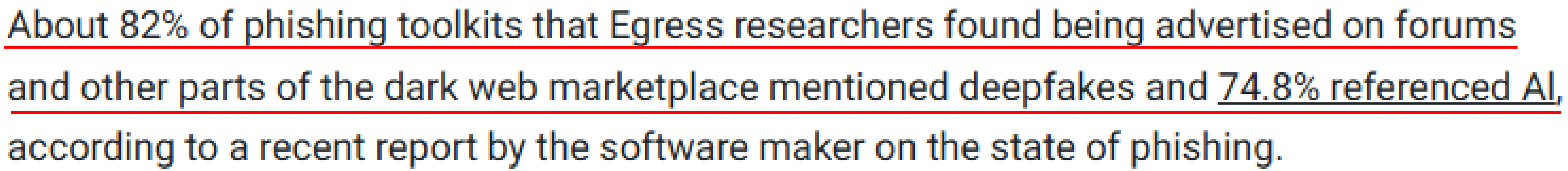
Among the other services advertised by the threat actor on a dedicated Telegram channel are AI-infused voice calling tools that allow its customers to generate voice calls to prospective targets based on a series of prompts directly from the phishing kit.



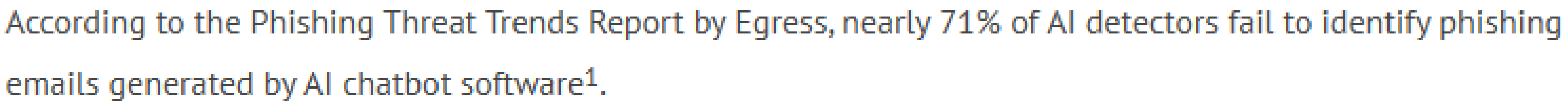
The phishing kit is priced anywhere between $150 and $900 a month,

https://thehackernews.com/2024/07/spanish-hackers-bundle-phishing-kits.html

AI-Enabled Malware Is Getting Popular



About 82% of phishing toolkits that Egress researchers found being advertised on forums and other parts of the dark web marketplace mentioned deepfakes and 74.8% referenced AI, according to a recent report by the software maker on the state of phishing.

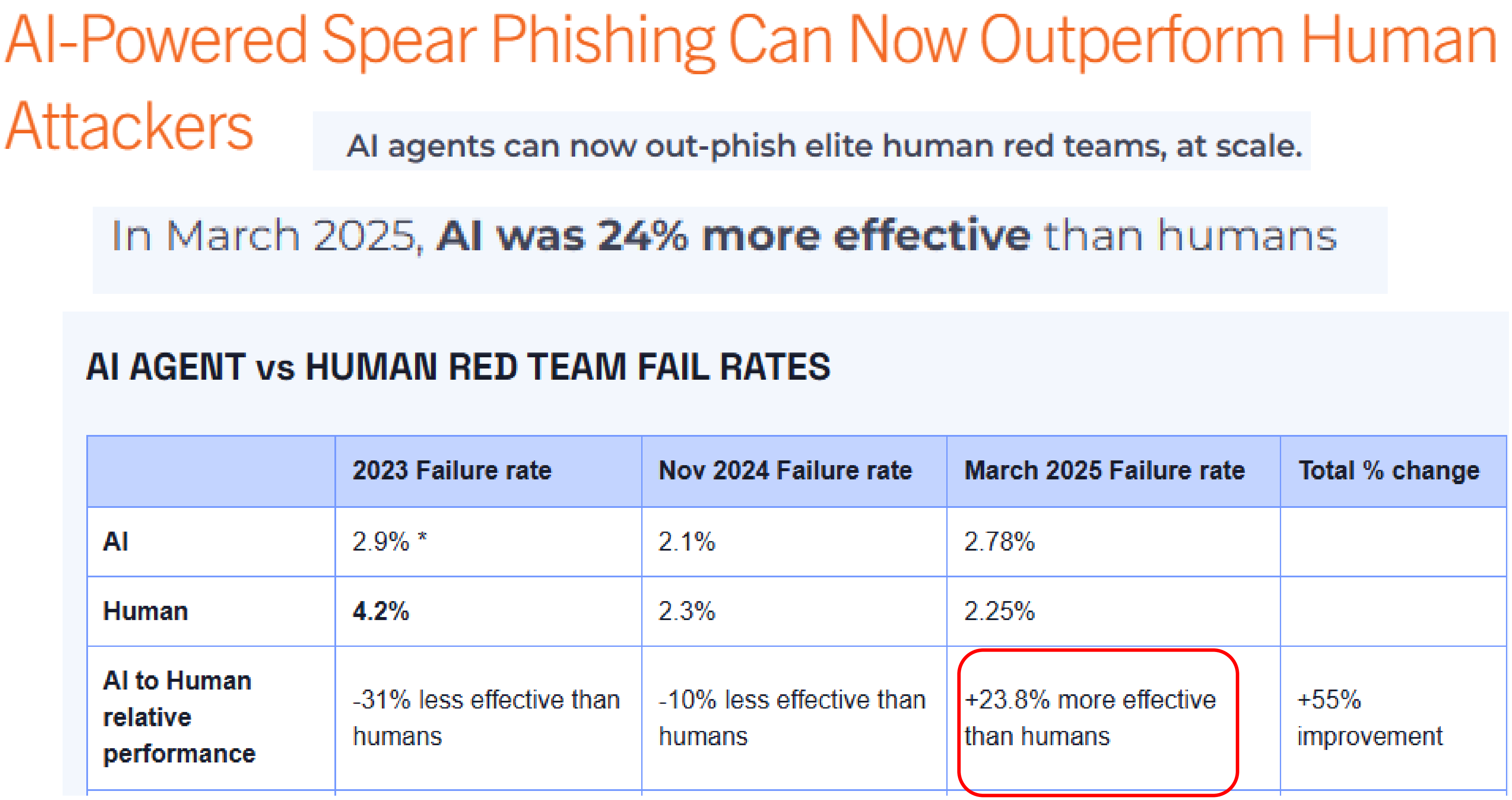


According to the Phishing Threat Trends Report by Egress, nearly 71% of AI detectors fail to identify phishing emails generated by AI chatbot software[1].

https://www.msspalert.com/analysis/ai-now-a-staple-in-phishing-kits-sold-to-hackers
https://www.ajg.com/uk/news-and-insights/the-rise-of-hyper-realistic-ai-powered-phishing/

Agentic AI Malware Is Getting Better

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



AI-Powered Spear Phishing Can Now Outperform Human Attackers

AI agents can now out-phish elite human red teams, at scale.

In March 2025, **AI was 24% more effective** than humans

AI AGENT vs HUMAN RED TEAM FAIL RATES

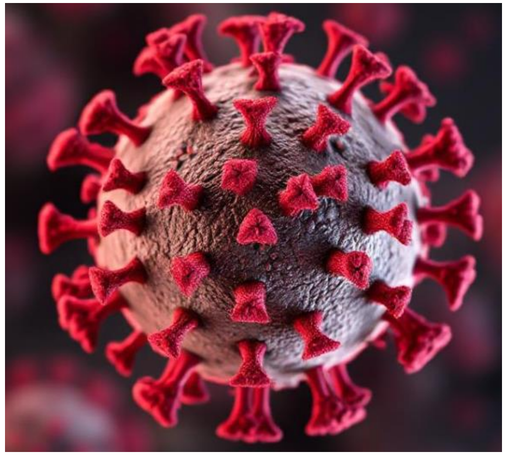
	2023 Failure rate	Nov 2024 Failure rate	March 2025 Failure rate	Total % change
AI	2.9% *	2.1%	2.78%	
Human	**4.2%**	2.3%	2.25%	
AI to Human relative performance	-31% less effective than humans	-10% less effective than humans	+23.8% more effective than humans	+55% improvement

https://blog.knowbe4.com/ai-powered-spear-phishing-can-now-outperform-human-attackers
https://hoxhunt.com/blog/ai-powered-phishing-vs-humans

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

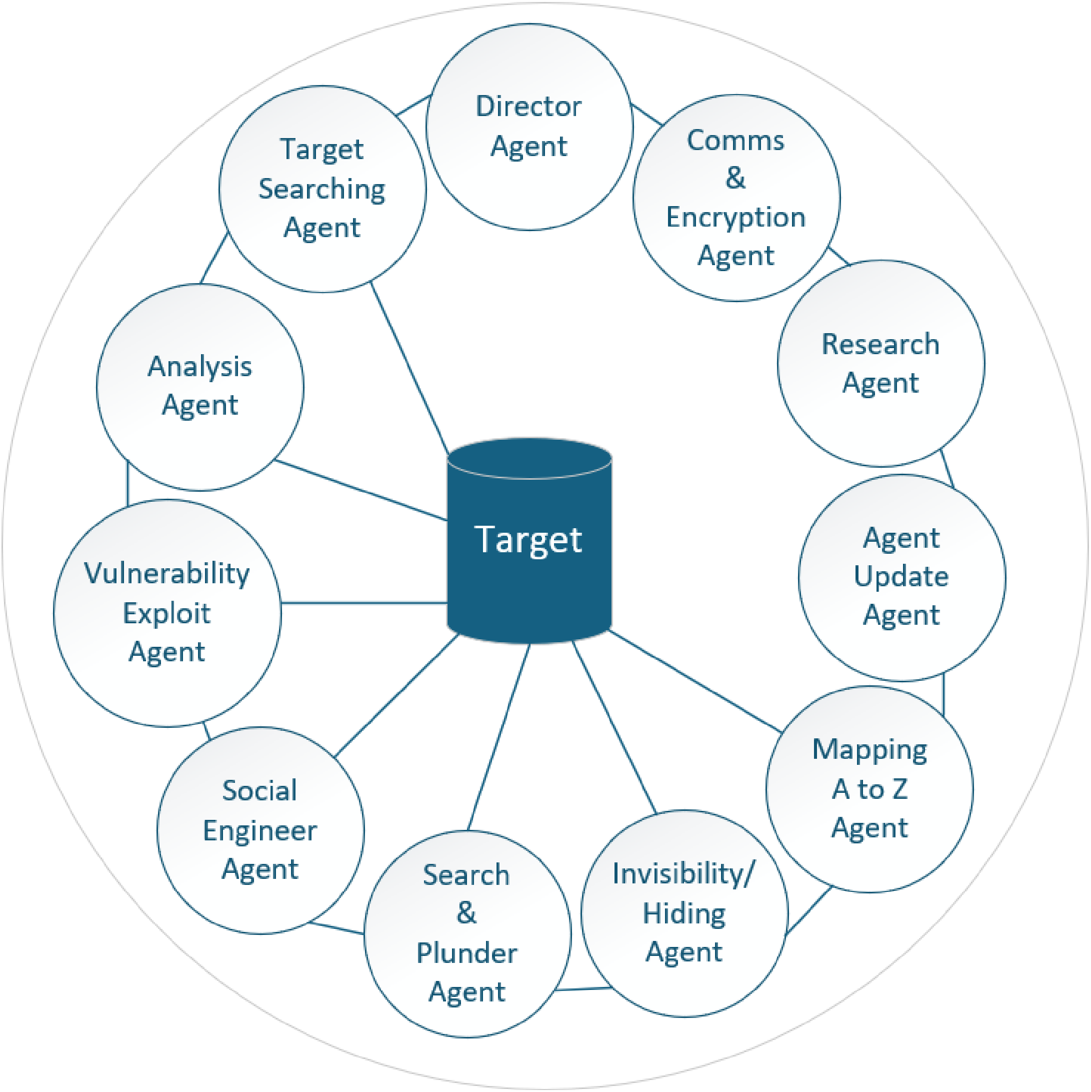


- This isn’t a possibility…this is what will happen

What Is Agentic AI Malware?

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

Potential Malicious Autonomous Agentic AI



Highly Autonomous Malware

Not Here Yet

But

It Will Be Soon

Social Engineering Agent

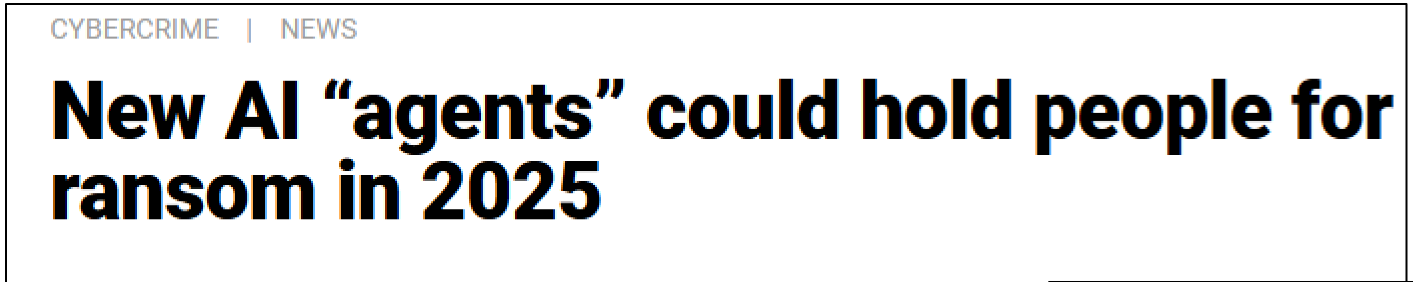
Less than 0.1% of email attacks are spearphishing, but are involved in 66% of all successful attacks – Barracuda Networks
https://www.barracuda.com/reports/spear-phishing-trends-2023

Agent Summary

It will customize each phishing attack for the exact victim

- It will research them on the Internet and socials
- Then construct a spearphishing attack with a high chance of success
- It will use your existing relationships (e.g., family, business, partners, group memberships, etc.) and interests against you

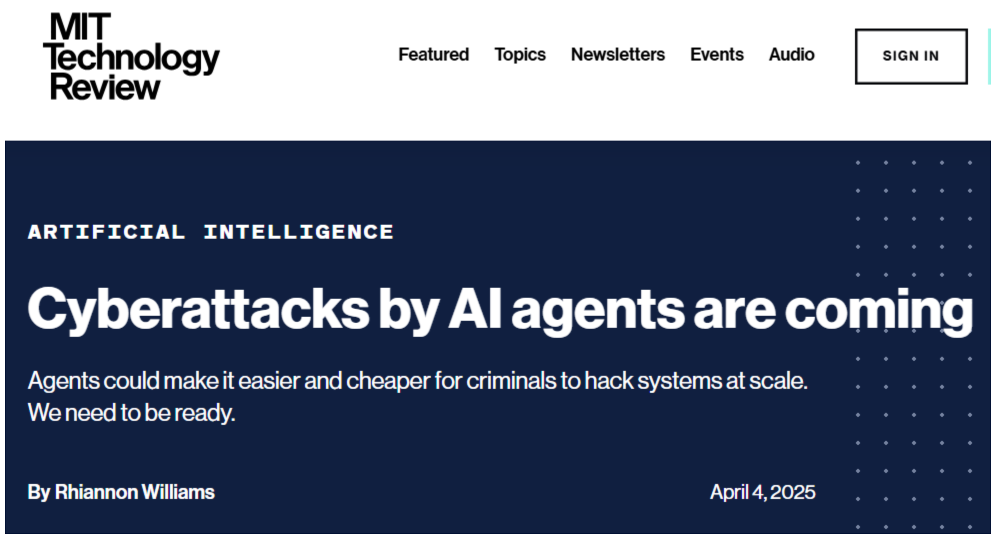
When Agentic AI Ransomware?



CYBERCRIME | NEWS

New AI "agents" could hold people for ransom in 2025

Posted: February 4, 2025 by David Ruiz



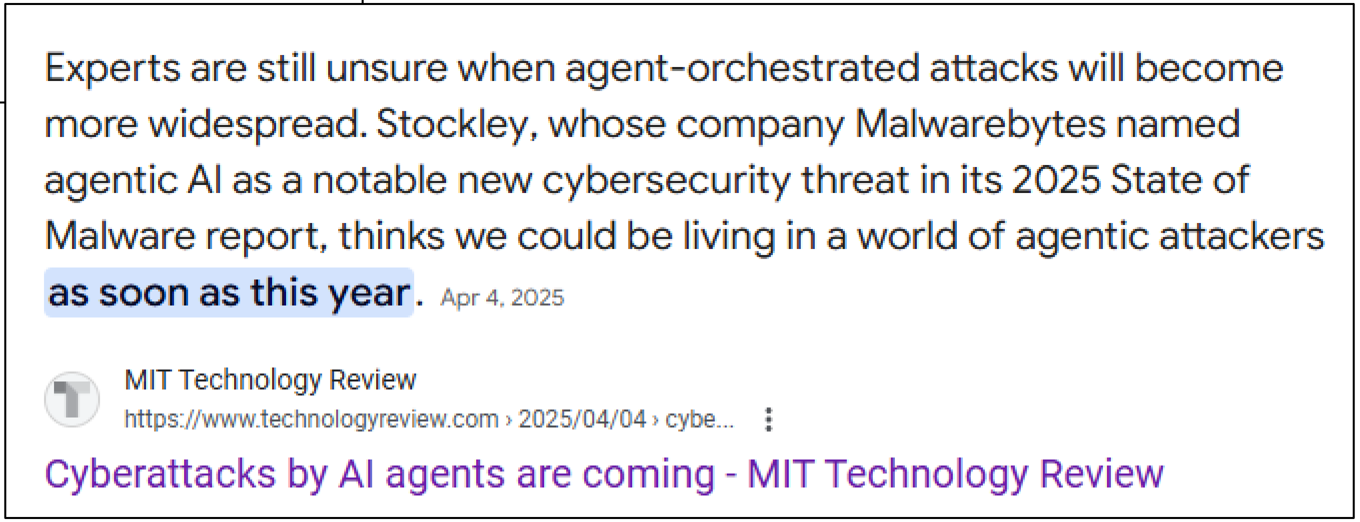
ARTIFICIAL INTELLIGENCE

Cyberattacks by AI agents are coming

Agents could make it easier and cheaper for criminals to hack systems at scale. We need to be ready.

By Rhiannon Williams

April 4, 2025



Experts are still unsure when agent-orchestrated attacks will become more widespread. Stockley, whose company Malwarebytes named agentic AI as a notable new cybersecurity threat in its 2025 State of Malware report, thinks we could be living in a world of agentic attackers as soon as this year. Apr 4, 2025

MIT Technology Review
https://www.technologyreview.com › 2025/04/04 › cybe...

Cyberattacks by AI agents are coming - MIT Technology Review

KnowBe4

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

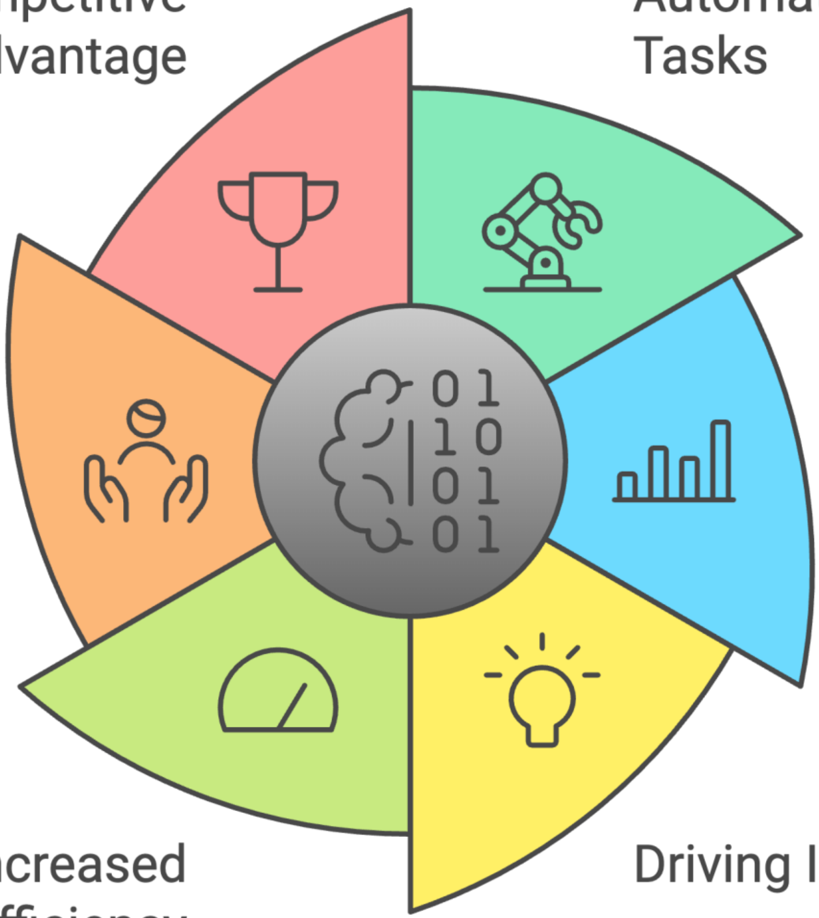
- This isn’t a possibility…this is what will happen

Agentic AI Defense Agents

- Inventory Agent
- Log Configuration/Analysis
- Authentication Analysis
- Cryptography Analysis
- Vulnerability Scanning
- Patch Management
- Pruning Agent
- Configuration Management
- Cybersecurity Training agents
- Network Traffic Analysis
- Malware Hunter
- Threat Hunting
- Anti-Denial-of-Service agents
- News/Research Agent
- Risk Management Analysis
- Deception Technologies
- Vendor Agentic AIs

Don’t Forget the Good Stuff

AI's Transformative Impact on Business



Competitive Advantage

Automation of Tasks

Improved Customer Experiences

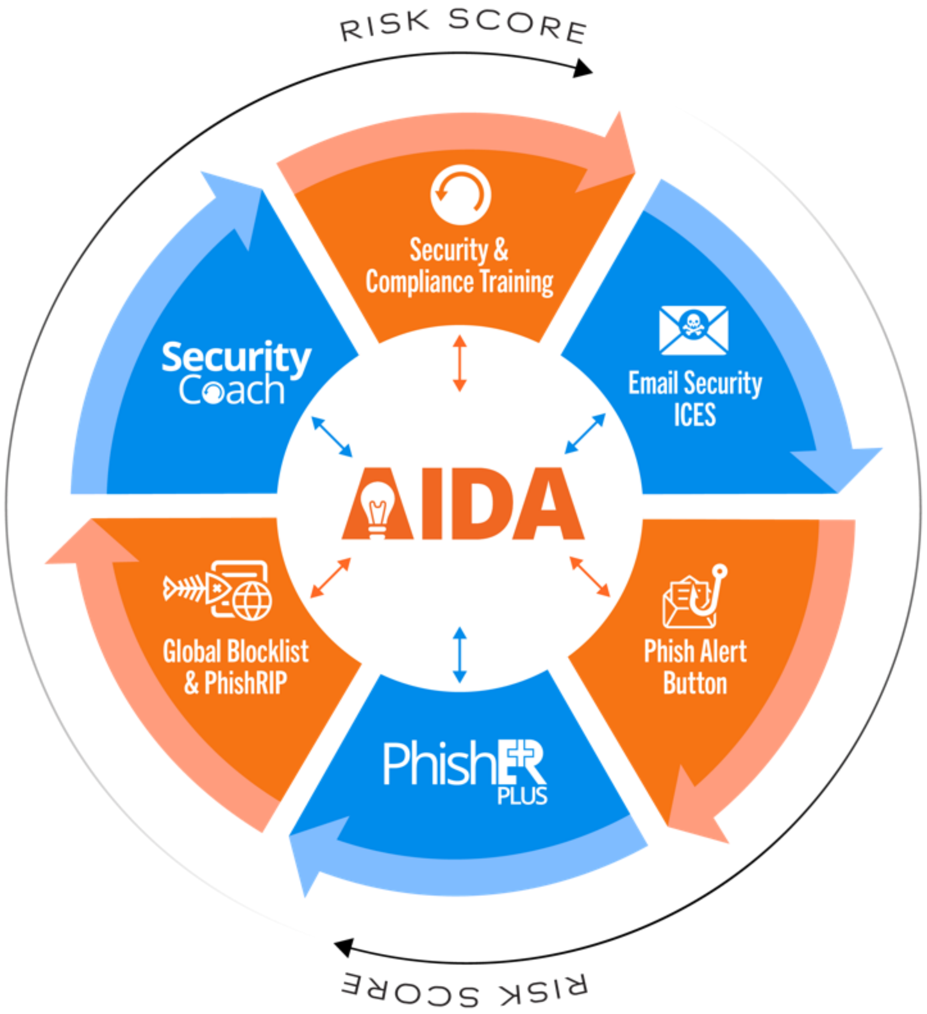
Enhanced Decision-Making

Increased Efficiency

Driving Innovation

How KnowBe4 Uses AI

AIDA



- AI-Driven Phishing
  - Unique Phishing Campaigns
  - Analyzes your users
  - Determines best templates for them
  - PAB-it
- AI-Recommended Optional Learning
  - Provide relevant training based on your users role
  - Create unique training for them

Defenses

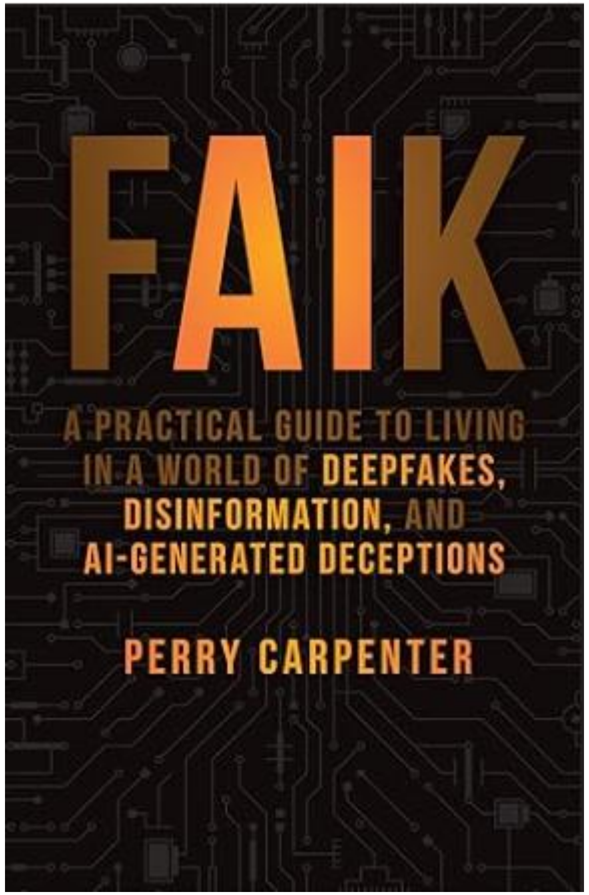
Overview

- Education, education, education

Best AI Defenses

To Prevent Malicious AI Problems

- Educate employees about malicious AI and deepfakes
- Help them detect, mitigate, and report AI-enabled attacks
- Focus on message, not typos
- Maybe use AI-enabled defense tools


FAIK
A PRACTICAL GUIDE TO LIVING IN A WORLD OF DEEPFAKES, DISINFORMATION, AND AI-GENERATED DECEPTIONS
PERRY CARPENTER

https://www.amazon.com/FAIK-Practical-Disinformation-AI-Generated-Deceptions/dp/1394299885

Best Defenses

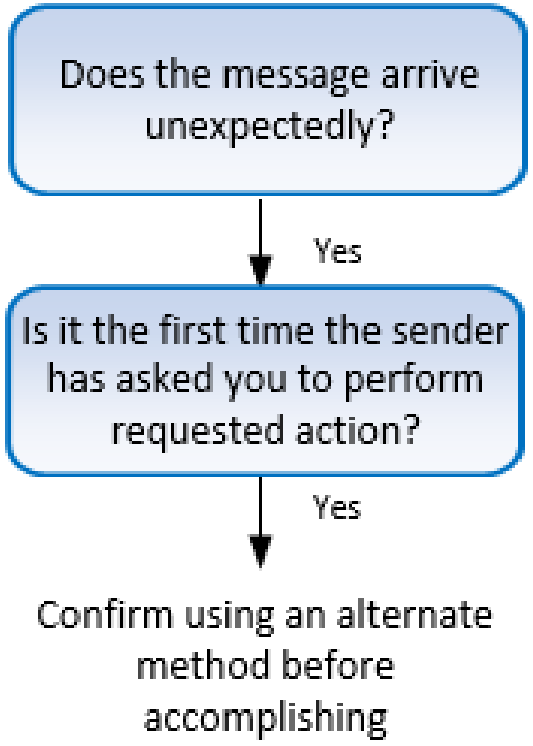
<u>Teach This!</u>

How

To

Spot

Phishing



Does the message arrive unexpectedly?

Yes

Is it the first time the sender has asked you to perform requested action?

Yes

Confirm using an alternate method before accomplishing

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

Best Way To Think About Deepfakes

Forget About Detection, Think About Motivation

- Nearly everything is going to have AI-enablement in the future, including nearly everything you do
- If true, AI-detection is really useless
- Instead of worrying about if it’s real or not, ask yourself if the message’s narrative could possibly have malicious intentions to motivate you to feel or act a certain way
- Take time to examine any message’s intention, especially if trying to emotionally motivate you to do something you otherwise wouldn’t do
- Try to prove or disprove any emotionally-motivating message, regardless of whether it’s AI-enabled or not, regardless of whether it is “real” or not.

https://blog.knowbe4.com/the-deceptive-media-era-moving-beyond-real-vs.-fake

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

KnowBe4

THANK YOU!

Roger A. Grimes– Data-Driven Defense Evangelist, KnowBe4

e: rogerg@knowbe4.com

LinkedIn: https://www.linkedin.com/in/rogeragrimes/

Mastodon: https://infosec.exchange/@rogeragrimes

Twitter: @RogerAGrimes

rogeragrimes@bsky.social