This presentation may contain simulated phishing attacks.

The trade names/trademarks of third parties used in this presentation are solely for illustrative and educational purposes.

The marks are property of their respective owners, and the use or display of the marks does not imply any affiliation with, endorsement by, or association of any kind between such third parties and KnowBe4.

Cybercriminals don't care about this and use them anyway to trick you….

Test Intro Video / Audio

KnowBe4

BSIDES 2024 KØBENHAVN

Digital Doppelganger

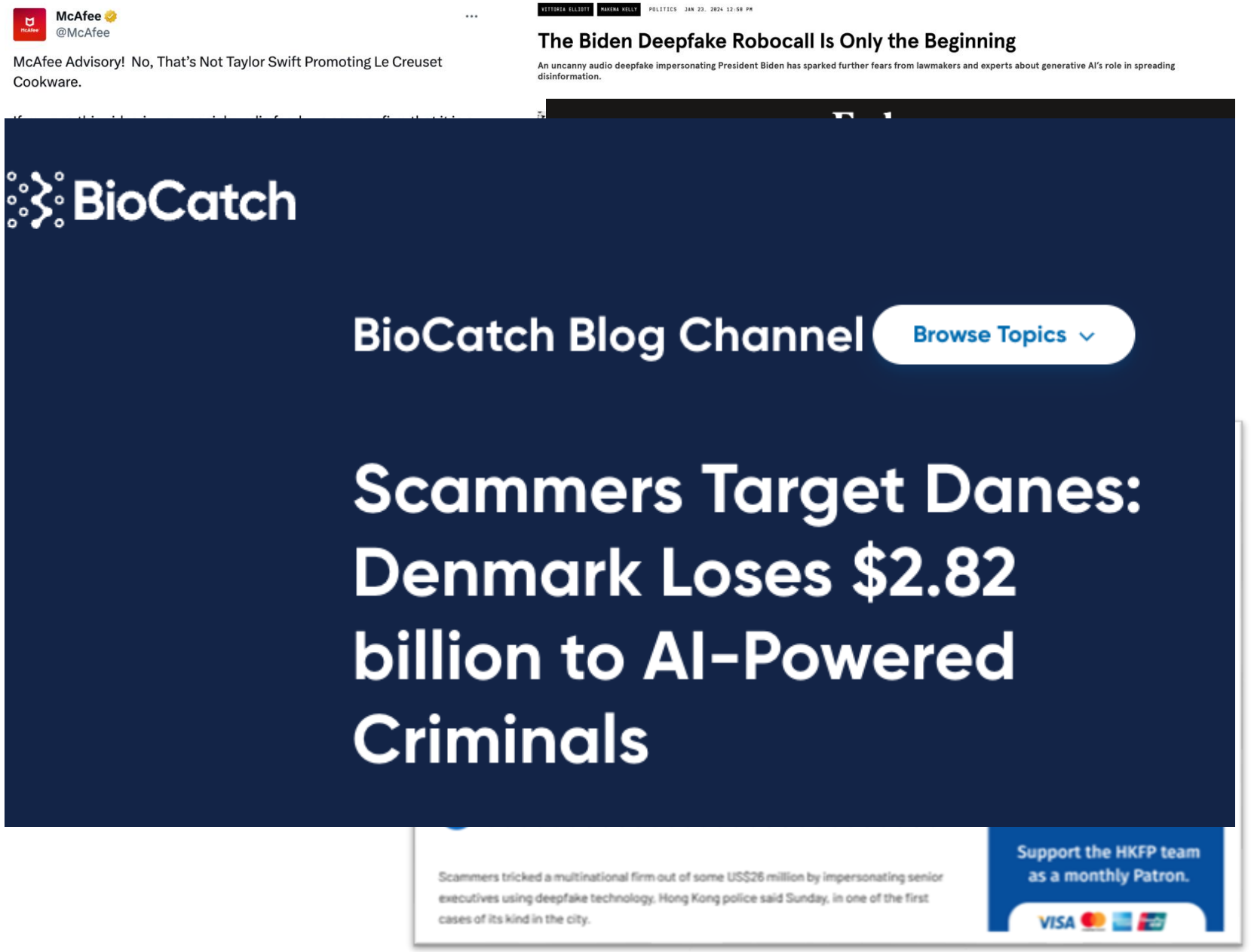
Deep fakes & Dark Side AI Attacks

James R. McQuiggan, CISSP, SACP
Security Awareness Advocate

Introduction

Synthetic media (deepfakes) are socially engineering our users to attack our organizations

Real World Synthetic Media (deepfakes) Attacks



How do we defend against synthetic media and protect our users and our organization?

James R. McQuiggan, CISSP,SACP,OSC

Security Awareness Advocate, KnowBe4 Inc.

Producer, Security Masterminds Podcast

Professor, Cyber Threat Intelligence, Full Sail

President, ISC2 Central Florida Chapter

ISC2 North American Advisory Council

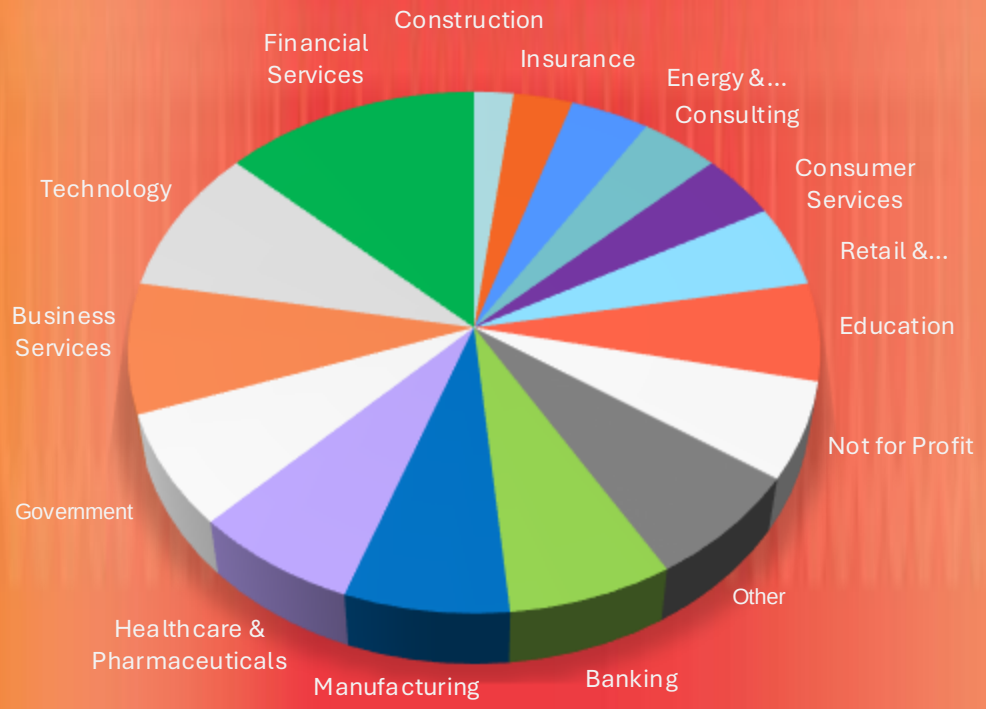
Cyber Security Awareness Lead, Siemens

Product Security Officer, Siemens Gamesa

Over

70,000

Customers



About KnowBe4

- The world's largest integrated Security Awareness Training and Simulated Phishing platform
- We help tens of thousands of organizations manage the ongoing problem of social engineering
- CEO & employees are industry veterans in IT Security
- Global Sales, Courseware Development, Customer Success, and Technical Support teams worldwide
- Offices in the USA, UK, Netherlands, India, Germany, South Africa, United Arab Emirates, Singapore, Japan, Australia, and Brazil

Our mission

To help organizations manage the ongoing problem of social engineering

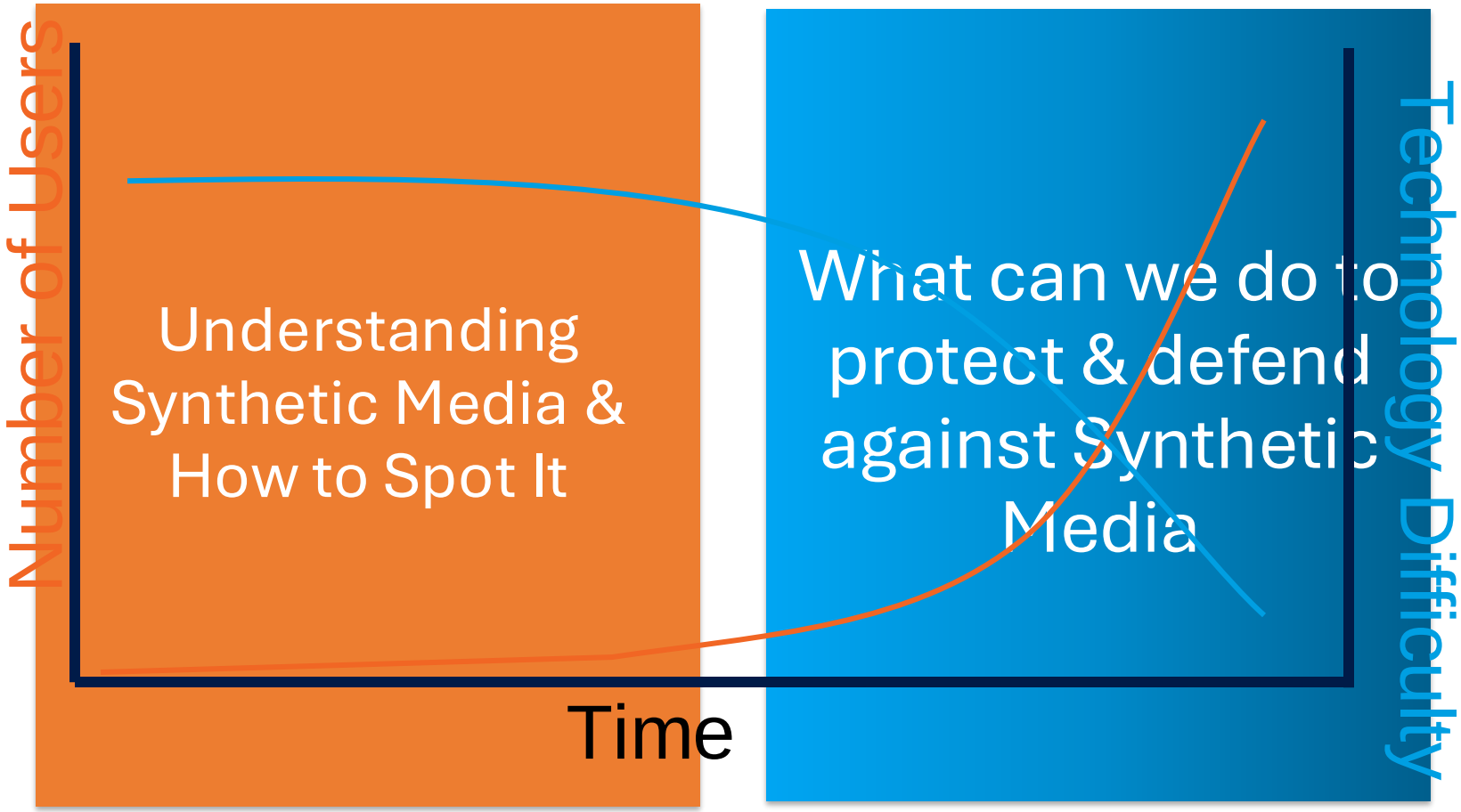
We do this by

Enabling employees to make smarter security decisions everyday


SECURITY
AWARENESS

Outcomes for the next 173 minutes… (and 375 slides)

Synthetic Media is quick and easy to create

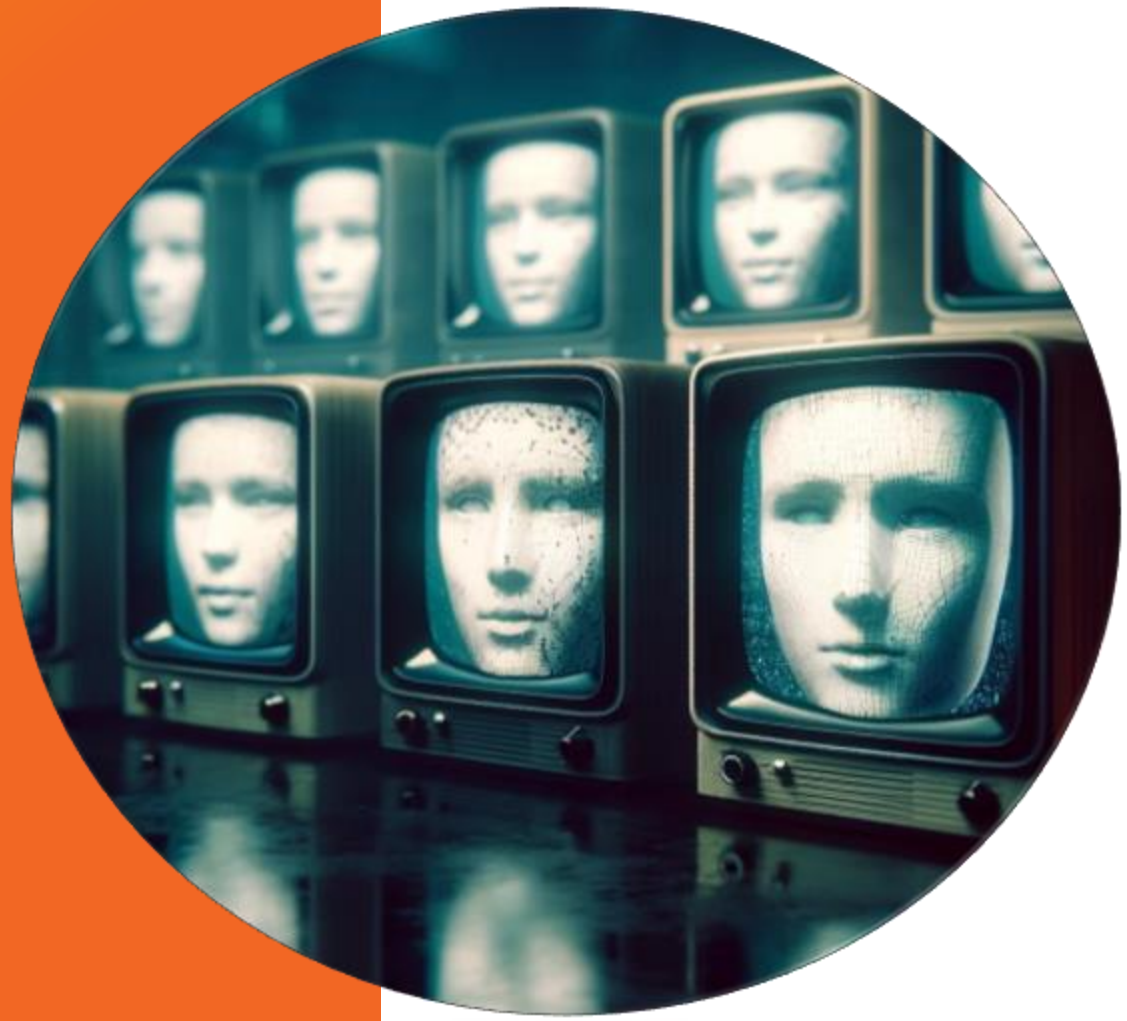


Synthetic Media

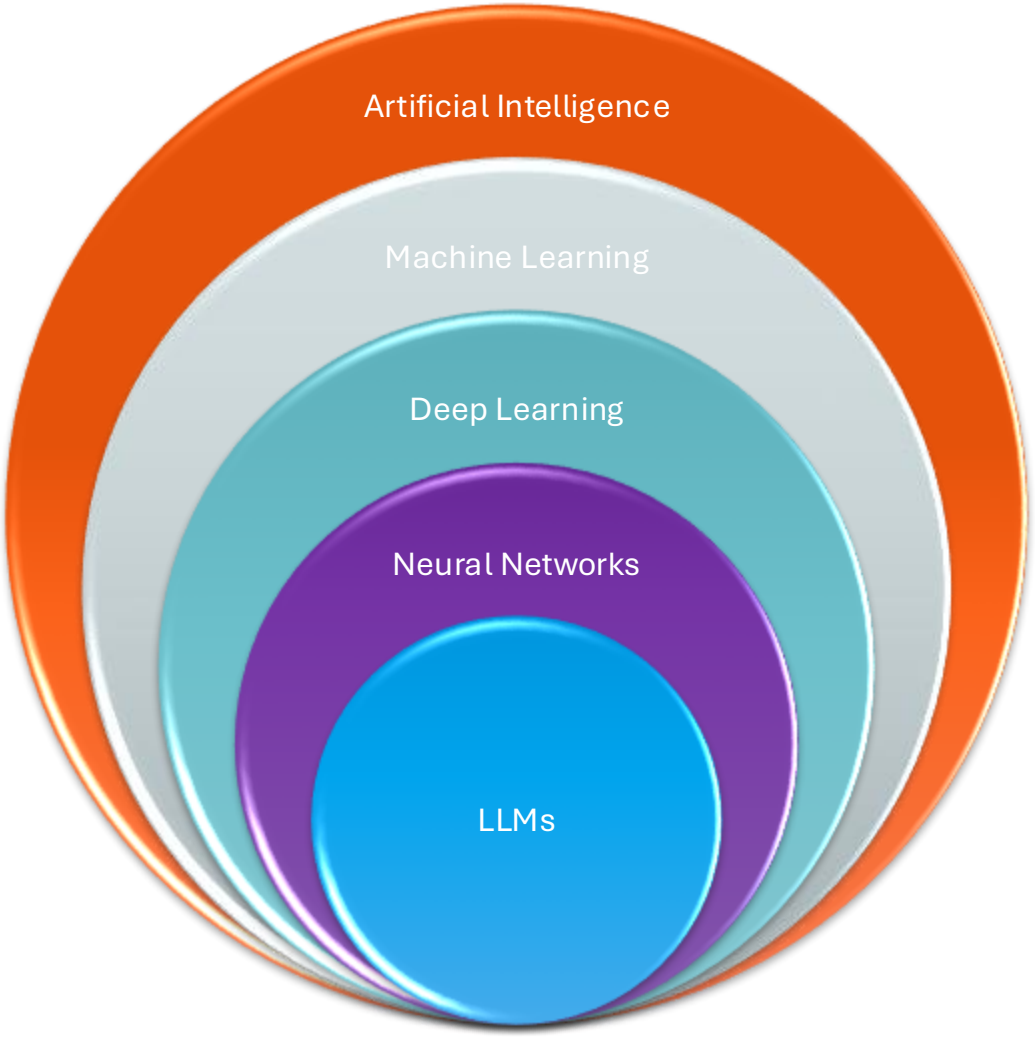
Detection / Prevention

Final Thoughts / Wrap-up

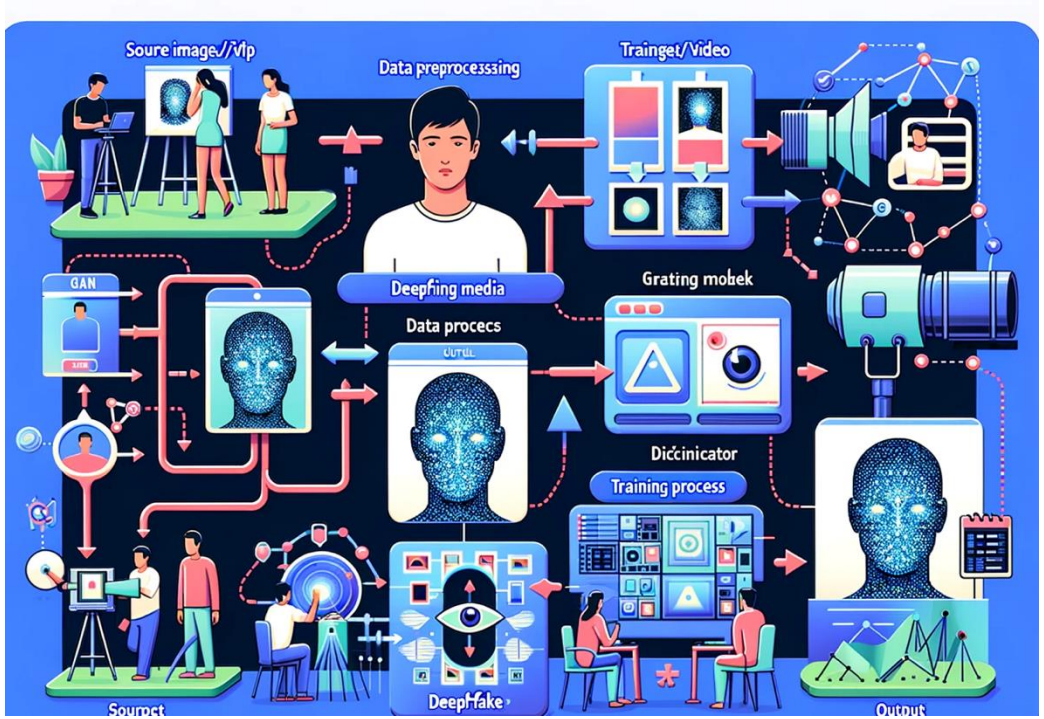
Synthetic Media



Artificial Intelligence & Synthetic Media

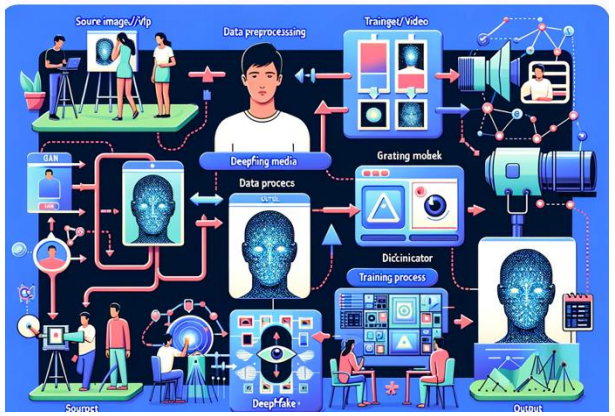


GAN Models



Generative Adversarial Models

GAN Models



Fools the discriminator

Identifies fakes

Creates synthetic data

Evaluates data realism

Generator

Discriminator

Roles and Goals of GAN Components

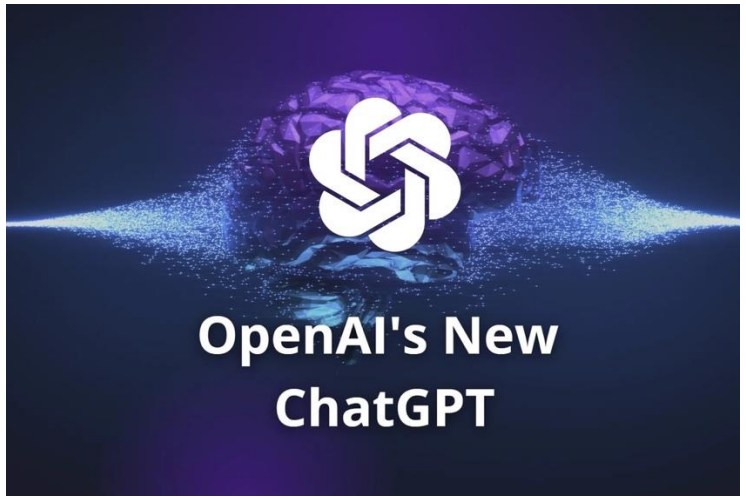
iminator

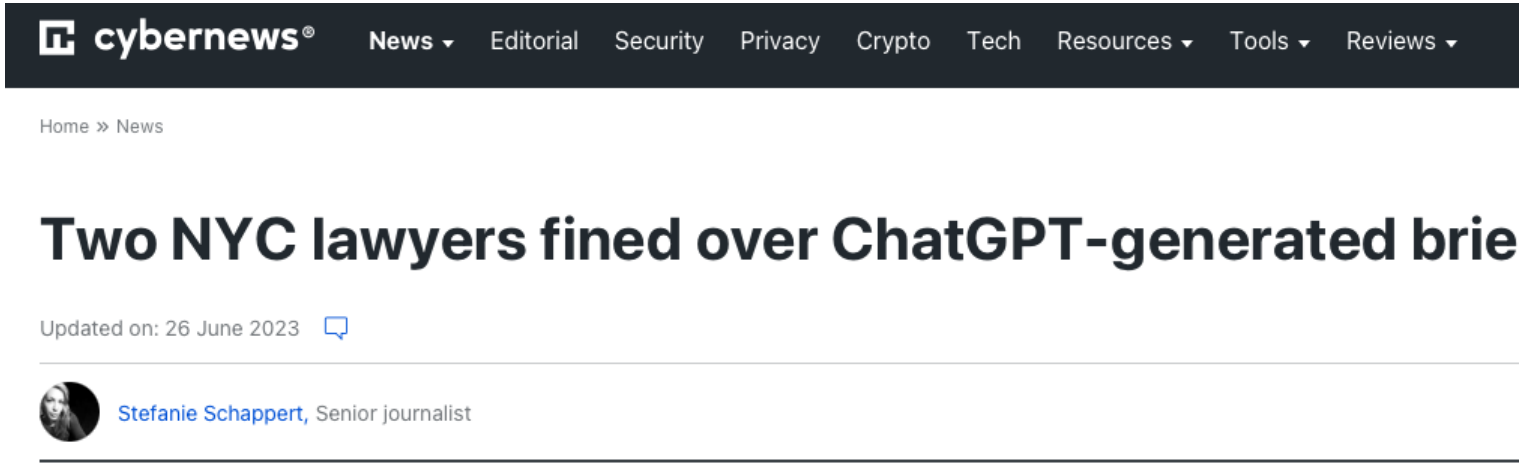
ata authenticity

Synthetic Text

Synthetic Text

- ChatGPT – exploded GenAI in November 2022
- Provided the ability to ask a question in the plain language and get a response back

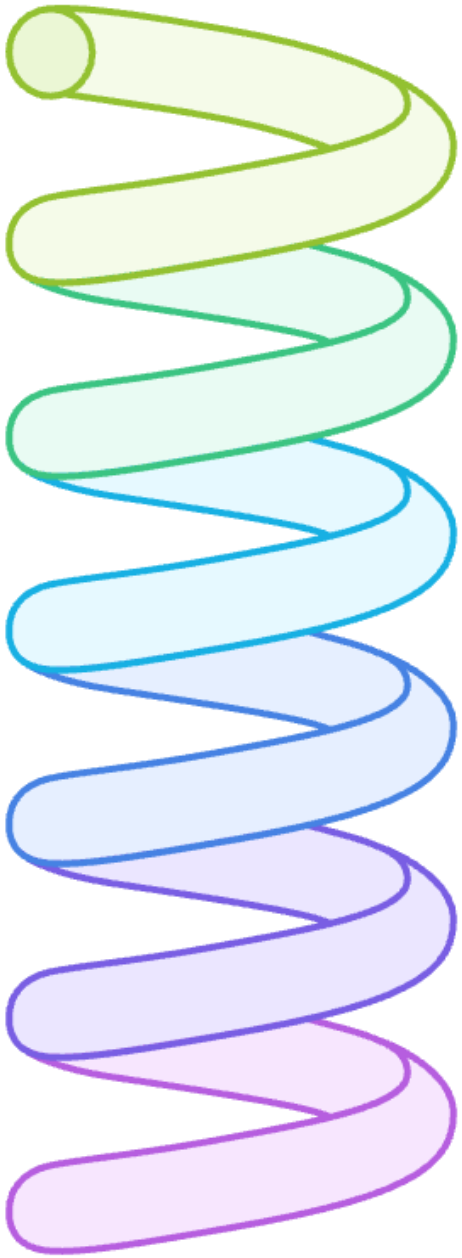




AI & Phishing

egress

a KnowBe4 company



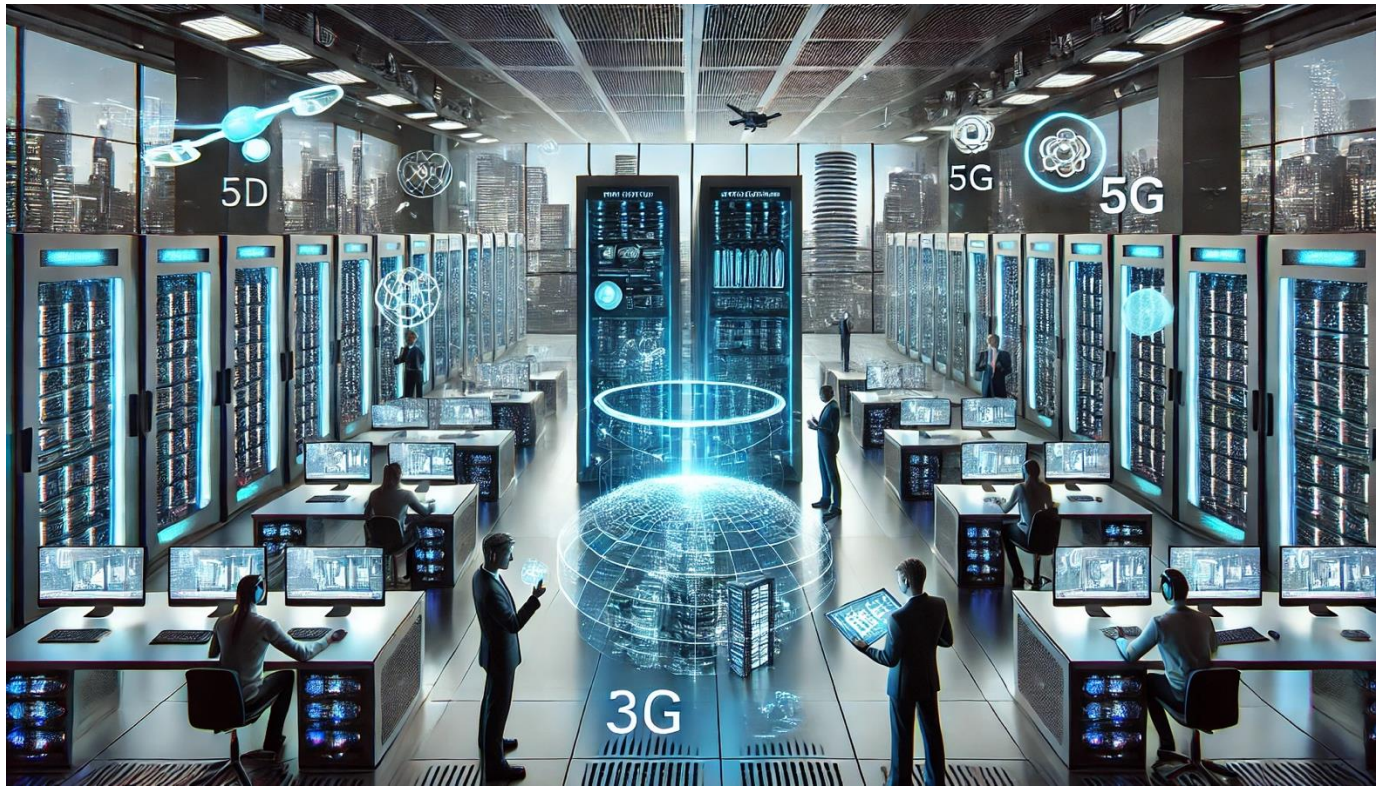
- Phishing attacks increase by 28%
- 44% of attacks originate from compromised accounts
- 45% of phishing emails contain malicious hyperlinks
- AI integrated into phishing toolkits
- 75% of kits offer AI features
- 82% of kits include deepfake capabilities

Source: https://www.scworld.com/news/phishing-attacks-armed-with-ai-capabilities-are-on-the-rise

Dark Web LLMs

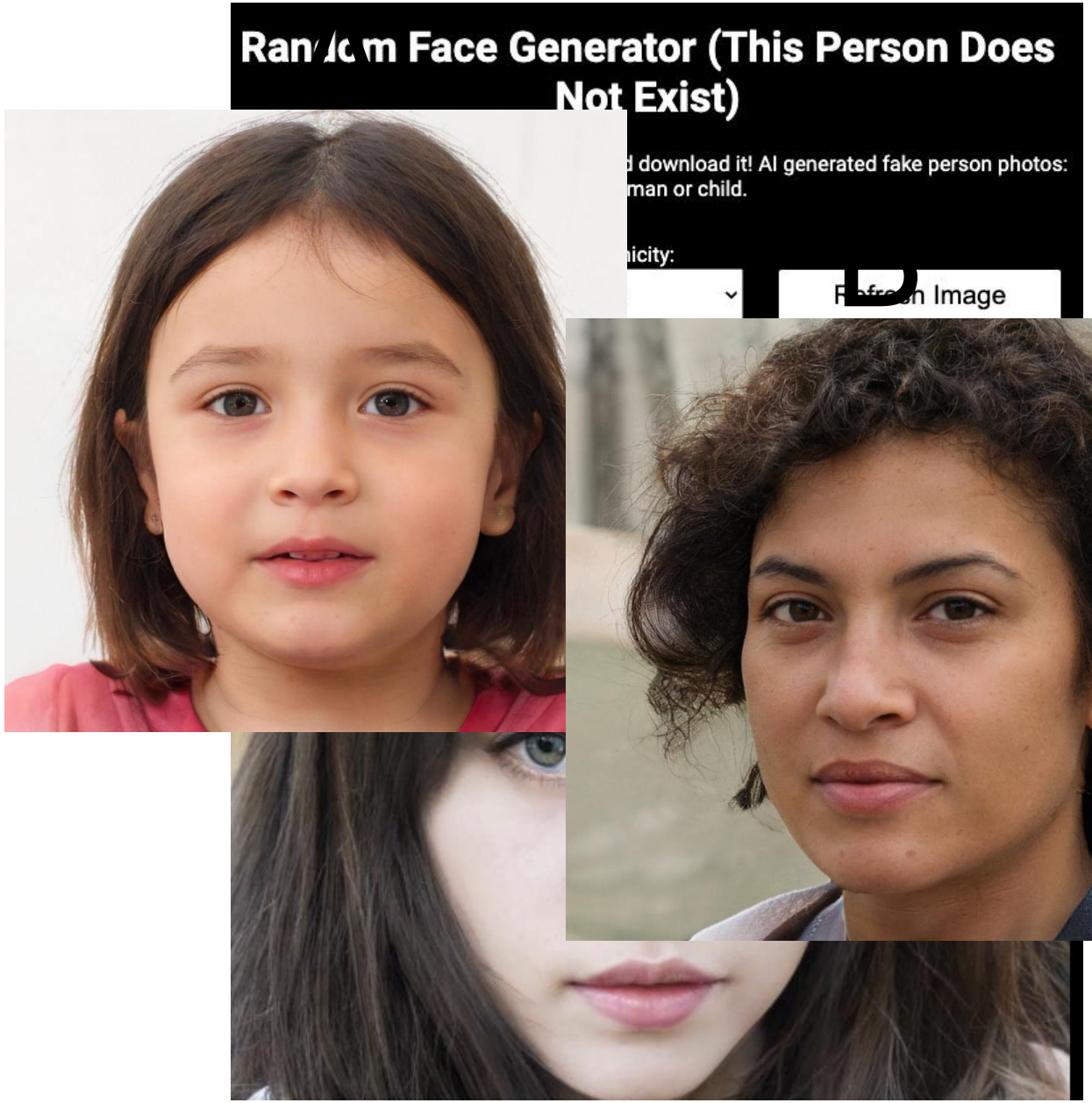
Synthetic Images

Text to Image – DALL-E, MidJourney

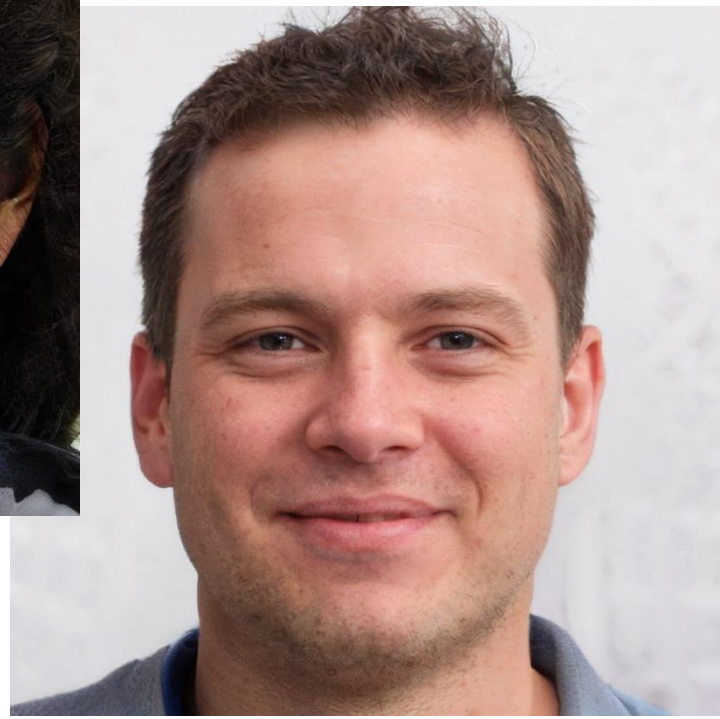


"Generate a futuristic data center scene with rows of sleek, glowing server racks connected by holographic networks. In the center, a highly advanced AI interface is projected in 3D, interacting with a group of IT professionals. Some professionals are analyzing data streams that float in mid-air, while others are remotely configuring servers through augmented reality glasses. The room is bathed in cool blue and white LED lighting, with advanced robotic systems assisting in managing the equipment. The background features a cityscape visible through glass windows, highlighting tall skyscrapers, drones flying, and 5G towers, emphasizing the advanced technological environment."

Synthetic Images



C

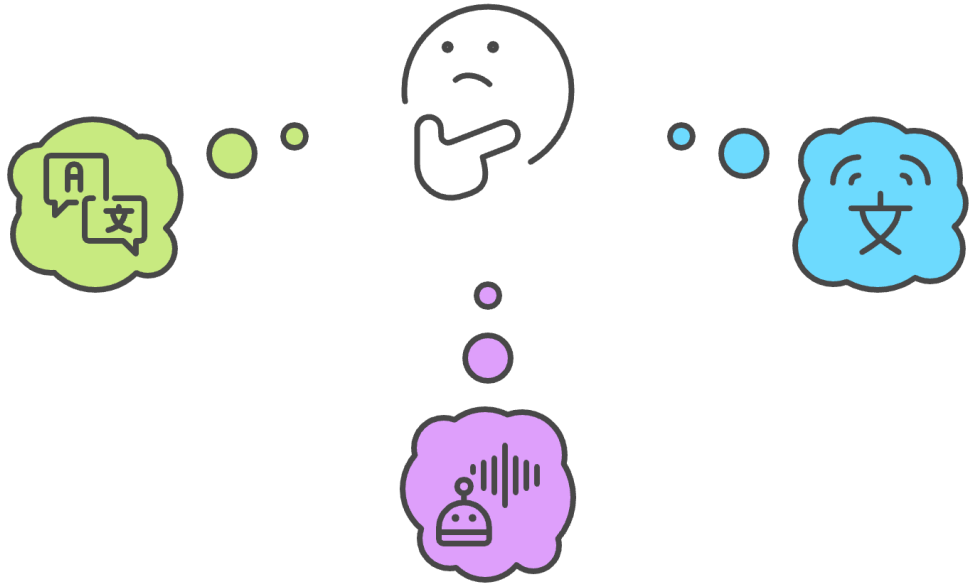


Synthetic Audio

Text to Audio

NLP Interpretation

Analyzes text for phonetics, syntax, tone, and inflection.



Speech Synthesis

Converts processed text into speech using deep learning models.

Voice Cloning

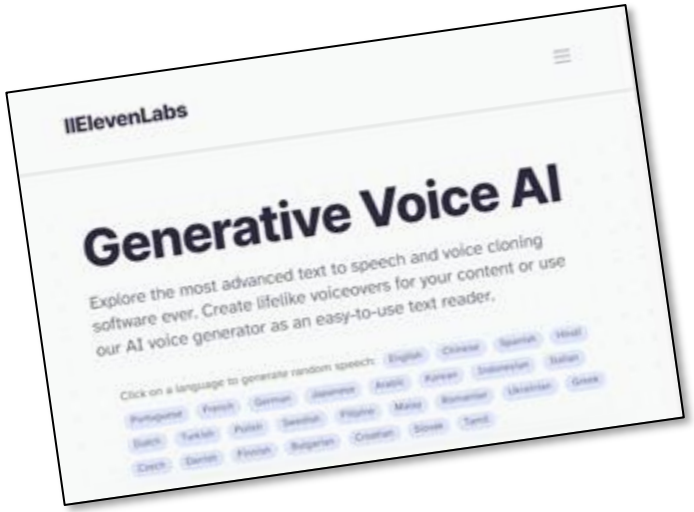
Matches generated speech with pre-trained voice models for naturalness.

Text to Audio

- Type the text you want
- Provide a voice
- AI Generates Audio

Common Tools

- 11Labs
- PlayHT
- Descript
- Resembler
- Respeecher



AI-Created Deepfakes Used In Attempt Theft

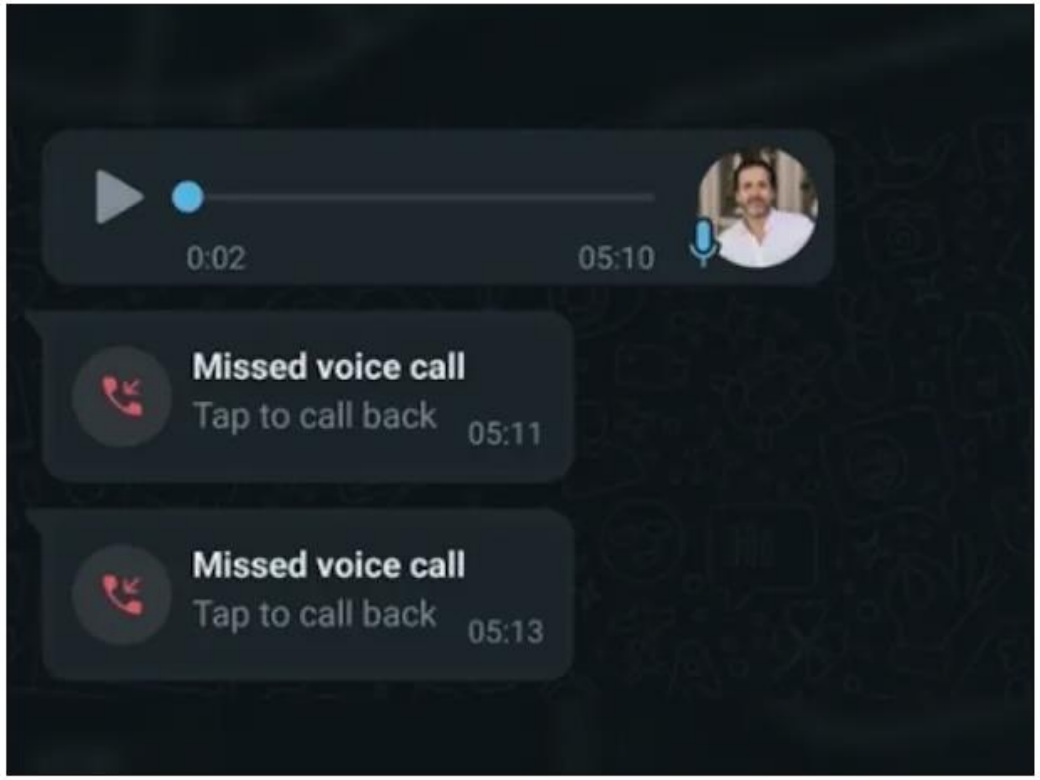
Fake Calls

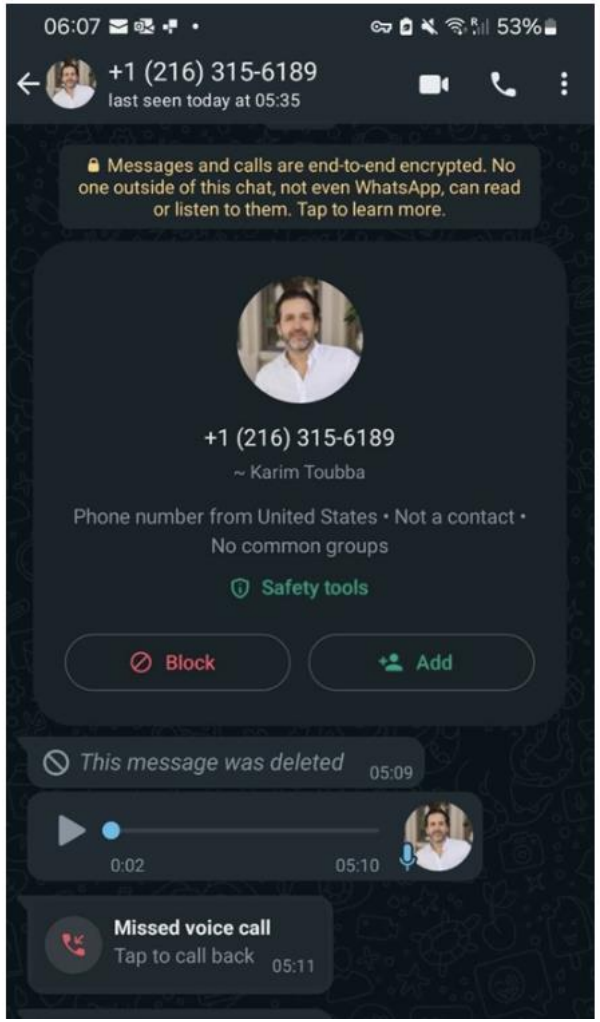
Audio Deepfake Attacks: Widespread And 'Only Going To Get Worse'

BY KYLE ALSPACH

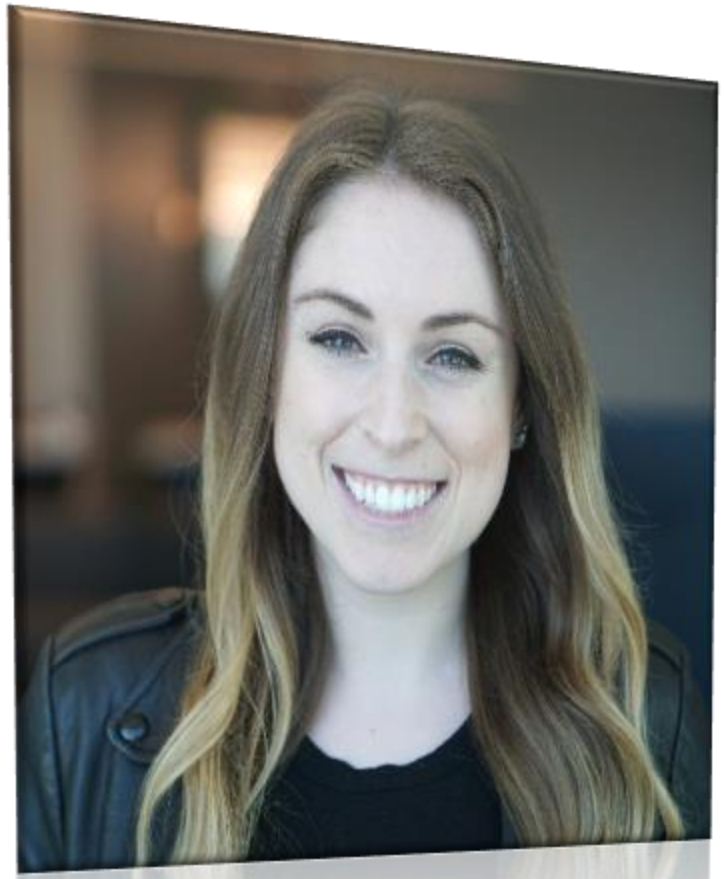
OCTOBER 3, 2024, 11:23 AM EDT

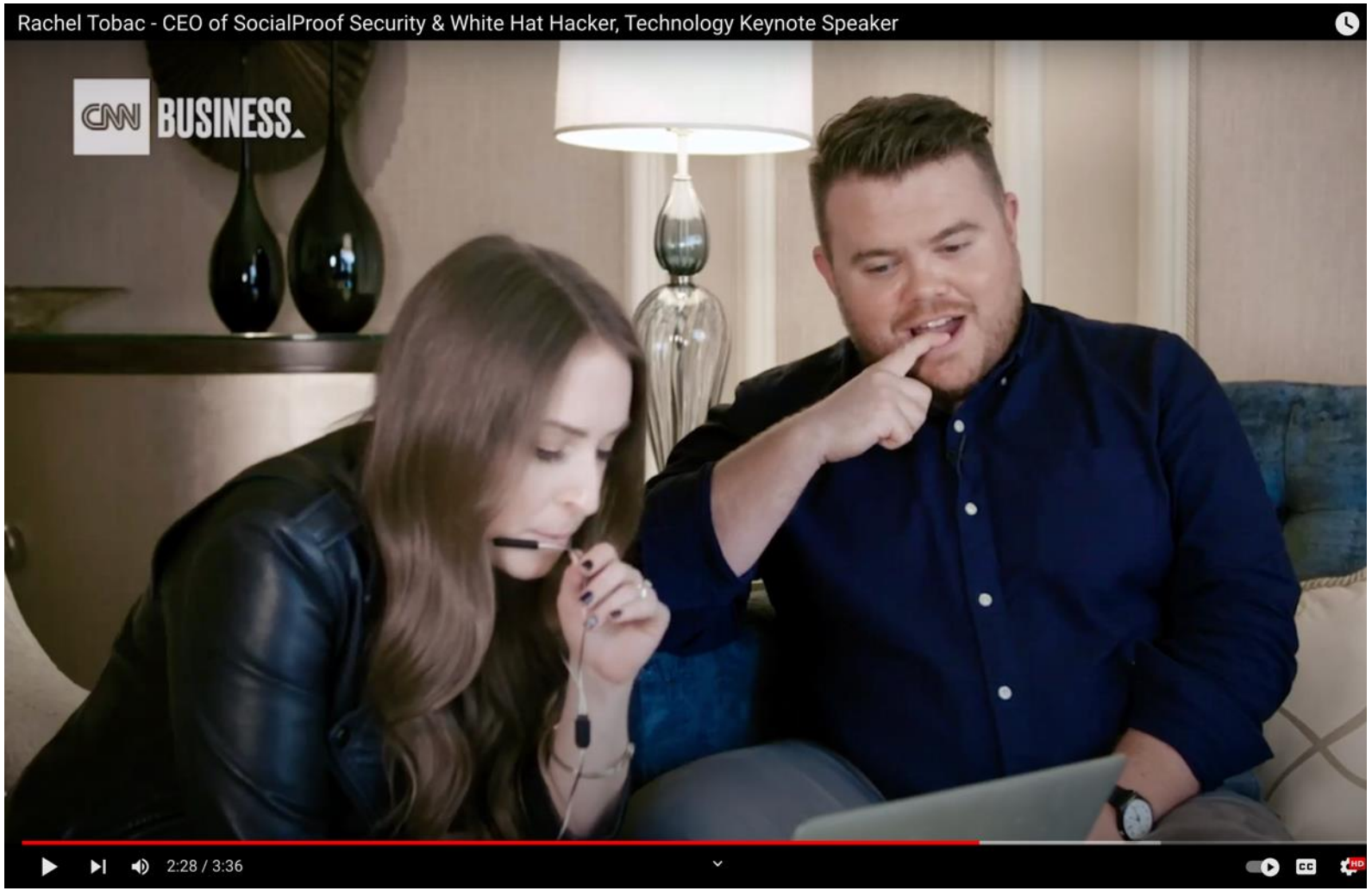
A cybersecurity researcher tells CRN that his own family was recently targeted with a convincing voice-clone scam.



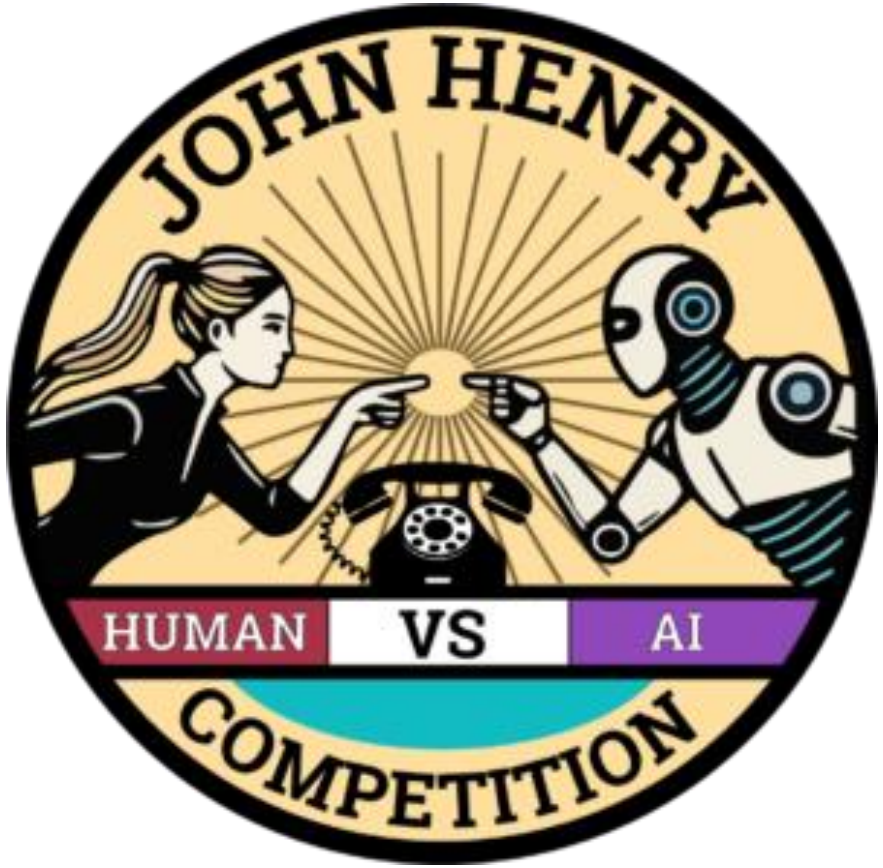


Audio Cloning – Rachel Tobac CNN





AI (ChatGPT + Syn Audio) = Conversation



SOCIAL ENGINEERING
COMMUNITY

Synthetic Audio & GenAI

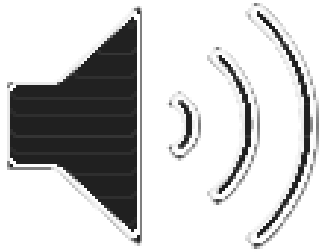
ChatGPT3.5 Prompt:

You are calling to tell me that you have been in a car accident and now he's being held by the police. Convince me that I need to send you $500 to pay the tow truck and start the repairs. You've been arrested and you don't have your wallet and you also need another $1500 to get you out of jail. The money needs to be sent as crypto currency as you know I have a crypto wallet and I can send money that way"

Using PlayHT

Dr. Gerald Auger, Simply Cyber (and friend)

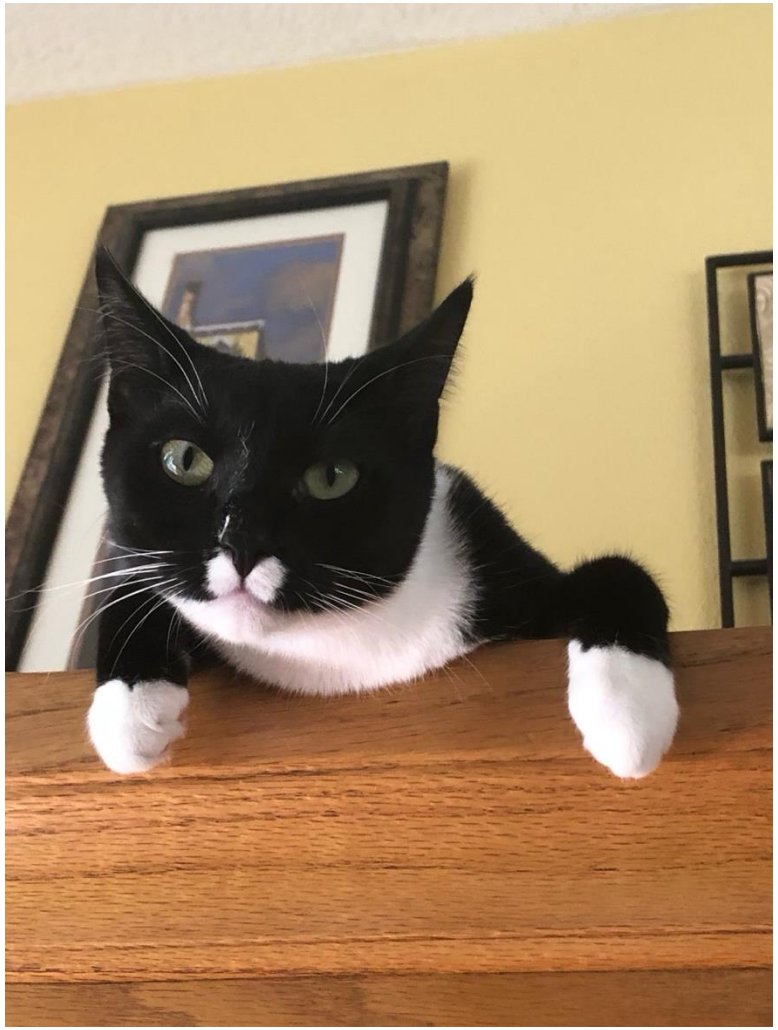
Call Center
Support
Software



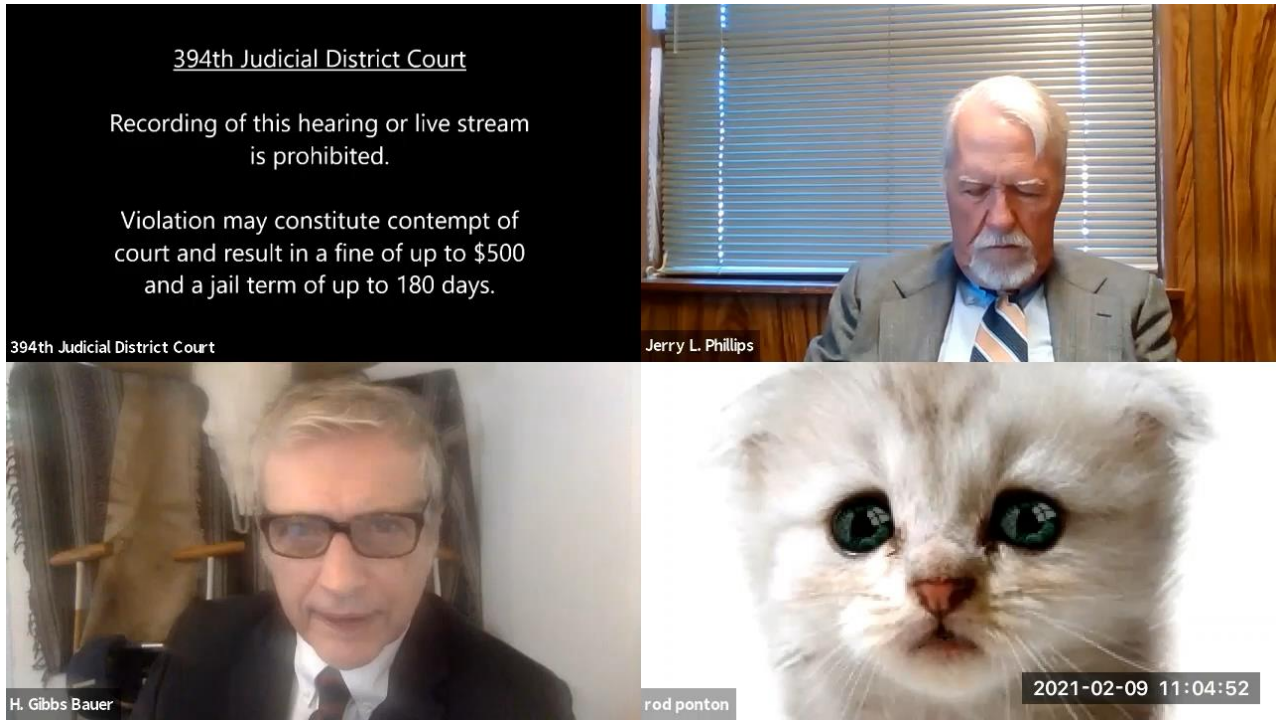
Synthetic Video

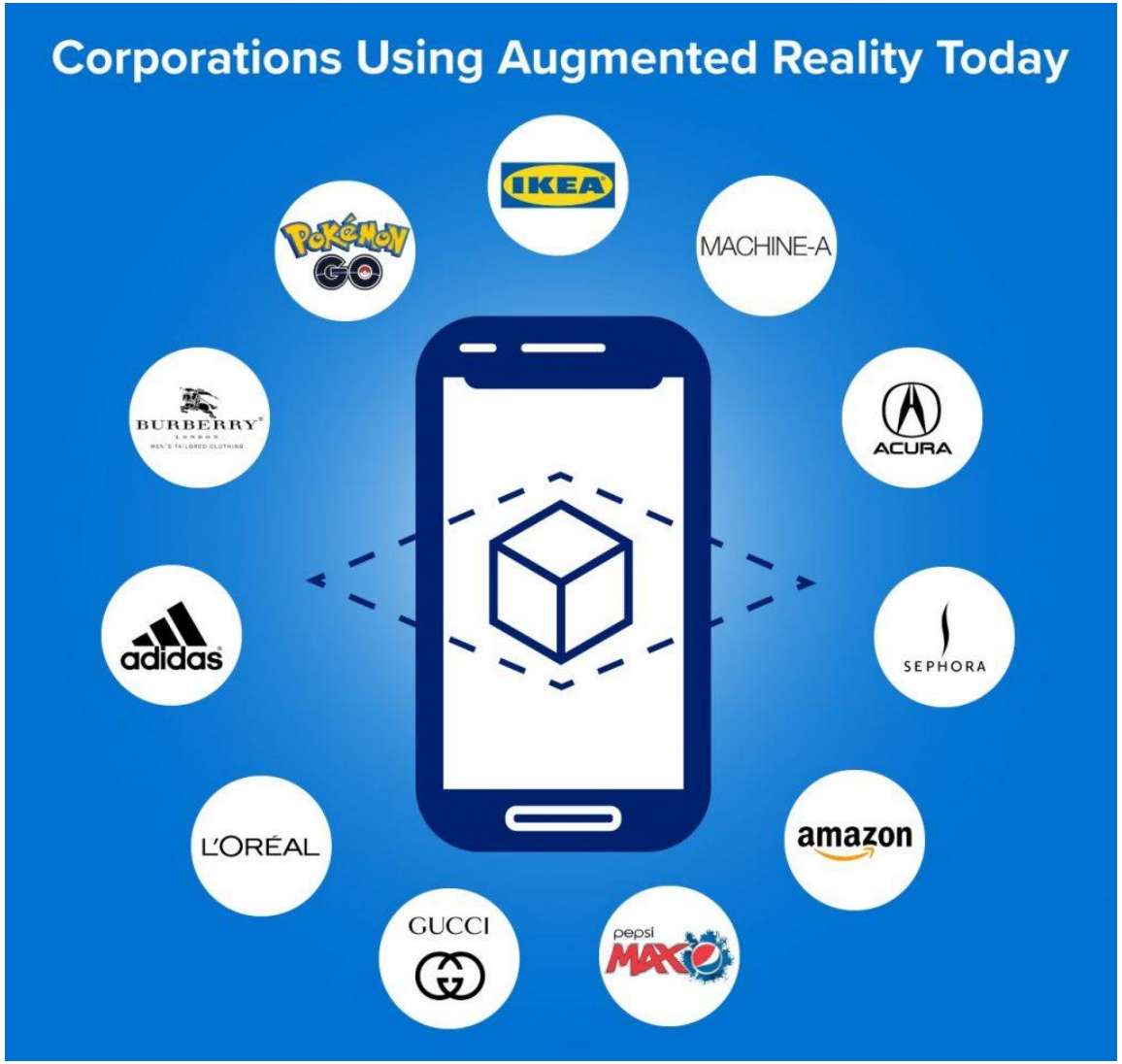
Modality – Image to Video

Modality – Image to Video



Augmented Reality





Deepfake Tools

Software

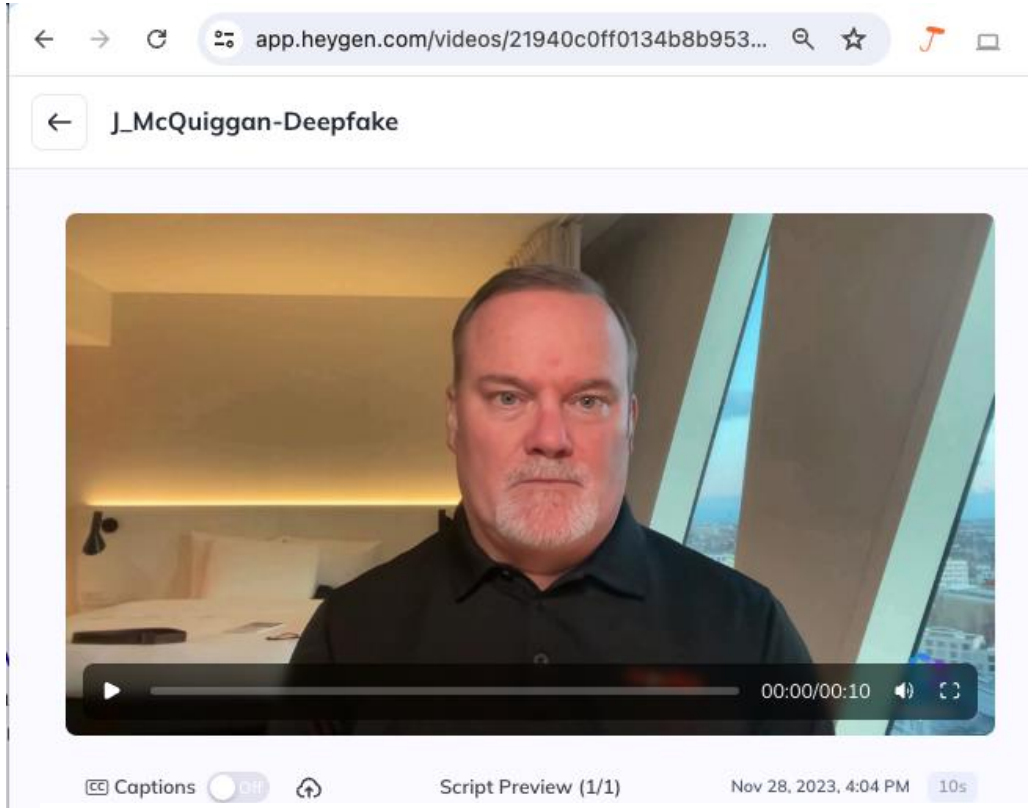
- Deepfakes Web
- Artguru AI
- Face Swap Live
- Reface
- UnDressVIP.com
- Lensa AI
- Wombo AI
- DeepSwap
- Synthesia

Platforms

- Smartphone apps
- SaaS
- Realtime / Stream
- Processed

DeepFaceLab - used in 95% of all deepfake videos

Modality – Text to Video (HeyGen Demo)

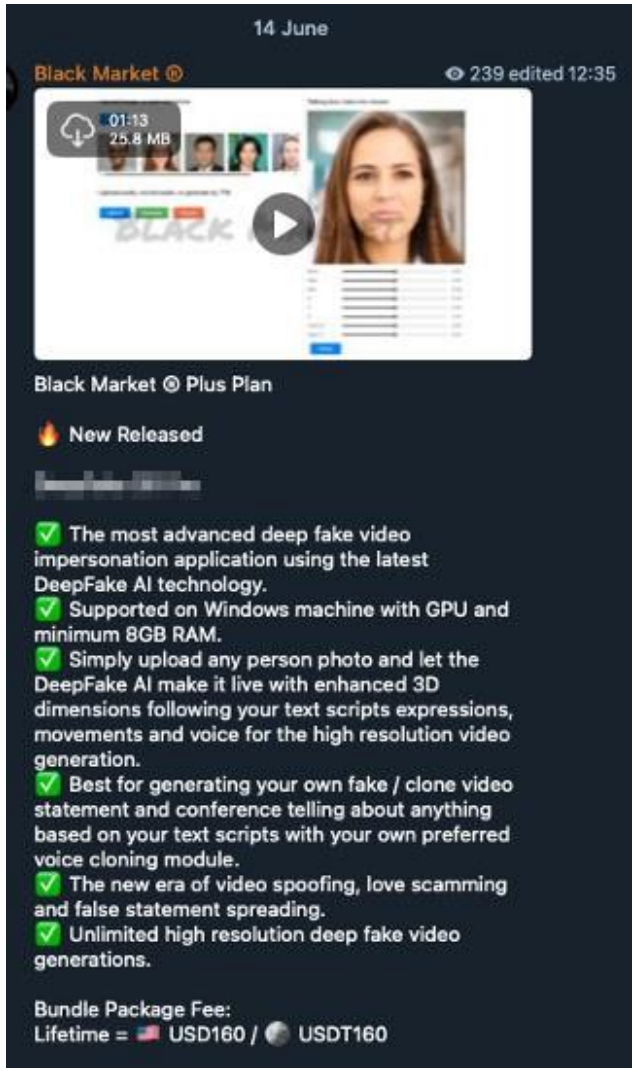


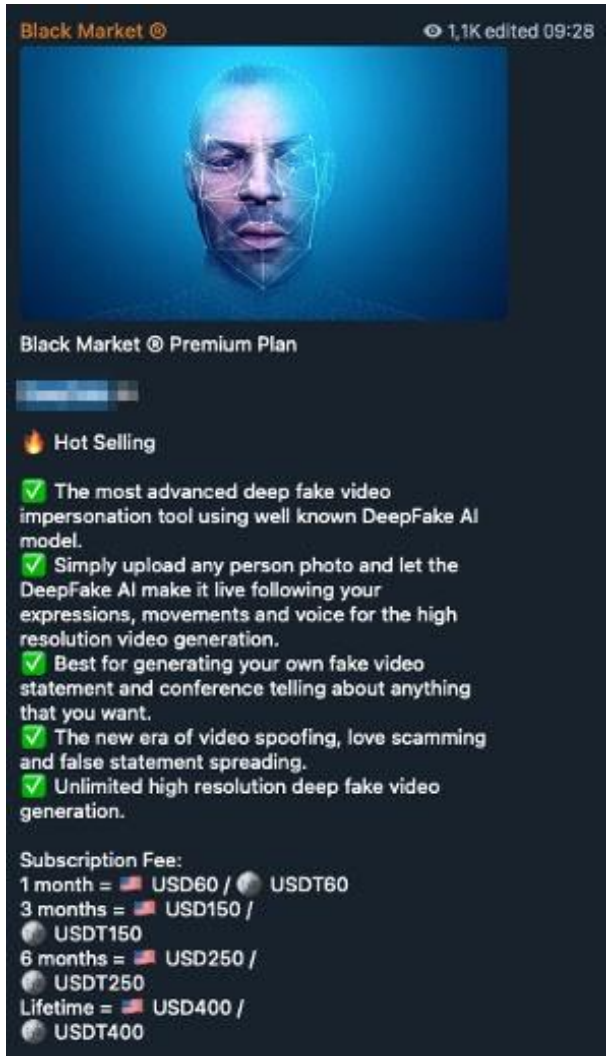
Now I took this one step further and created this audio and video which I recorded from a hotel room using HeyGen and all I did was type in a script of what I wanted to say.

Creating The Deepfake

- YouTube channel – pulled video / audio
- Imported to 11Labs to generate Audio file
- Imported Headshot and audio from 11Labs to generate video on Hedra
- Good thing I asked Andrada for permission!

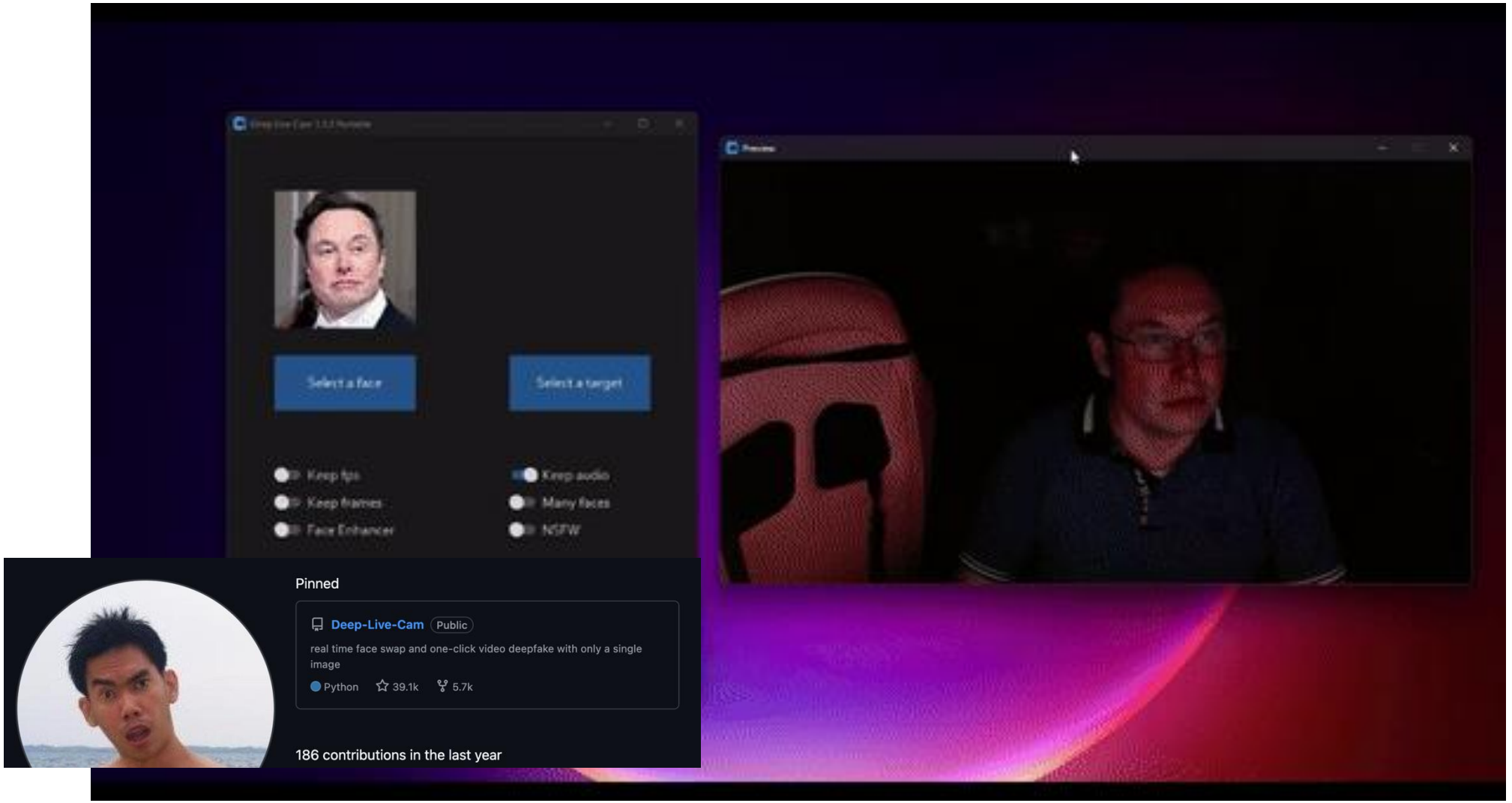
Dark Web Activity





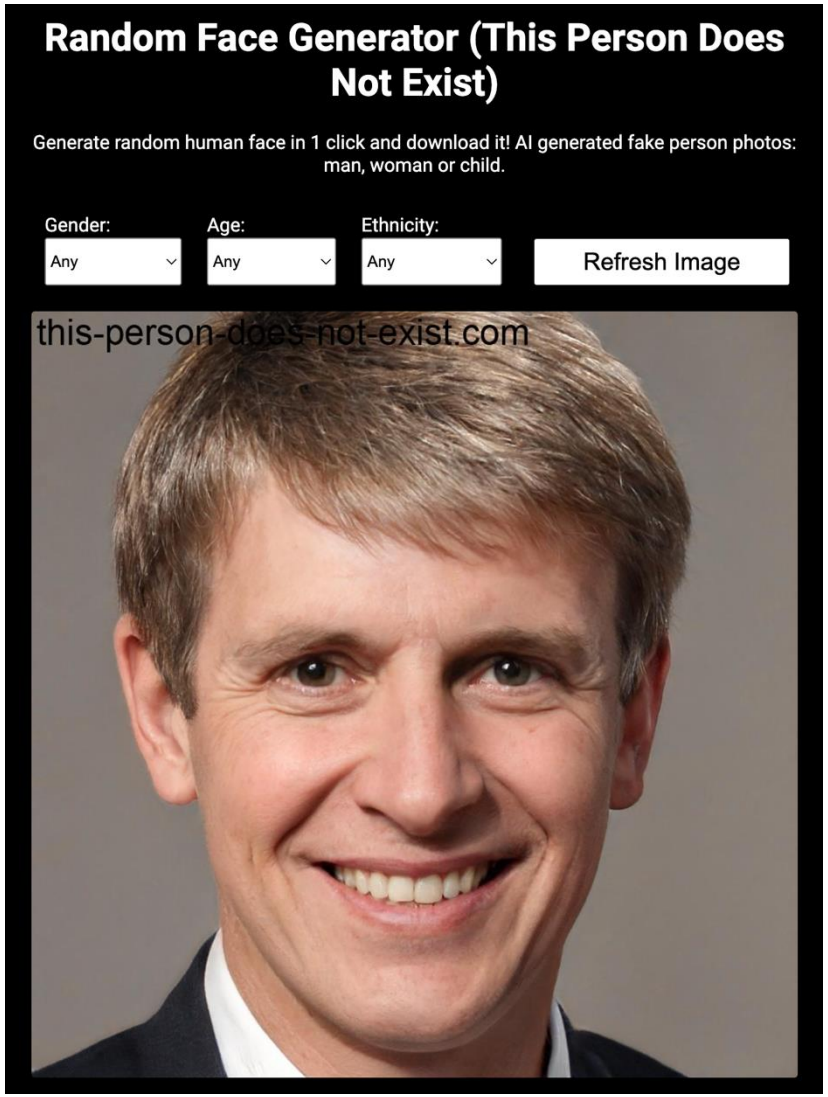
Source: https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/surging-hype-an-update-on-the-rising-abuse-of-genai

Deepfake & Webcams - LIVE

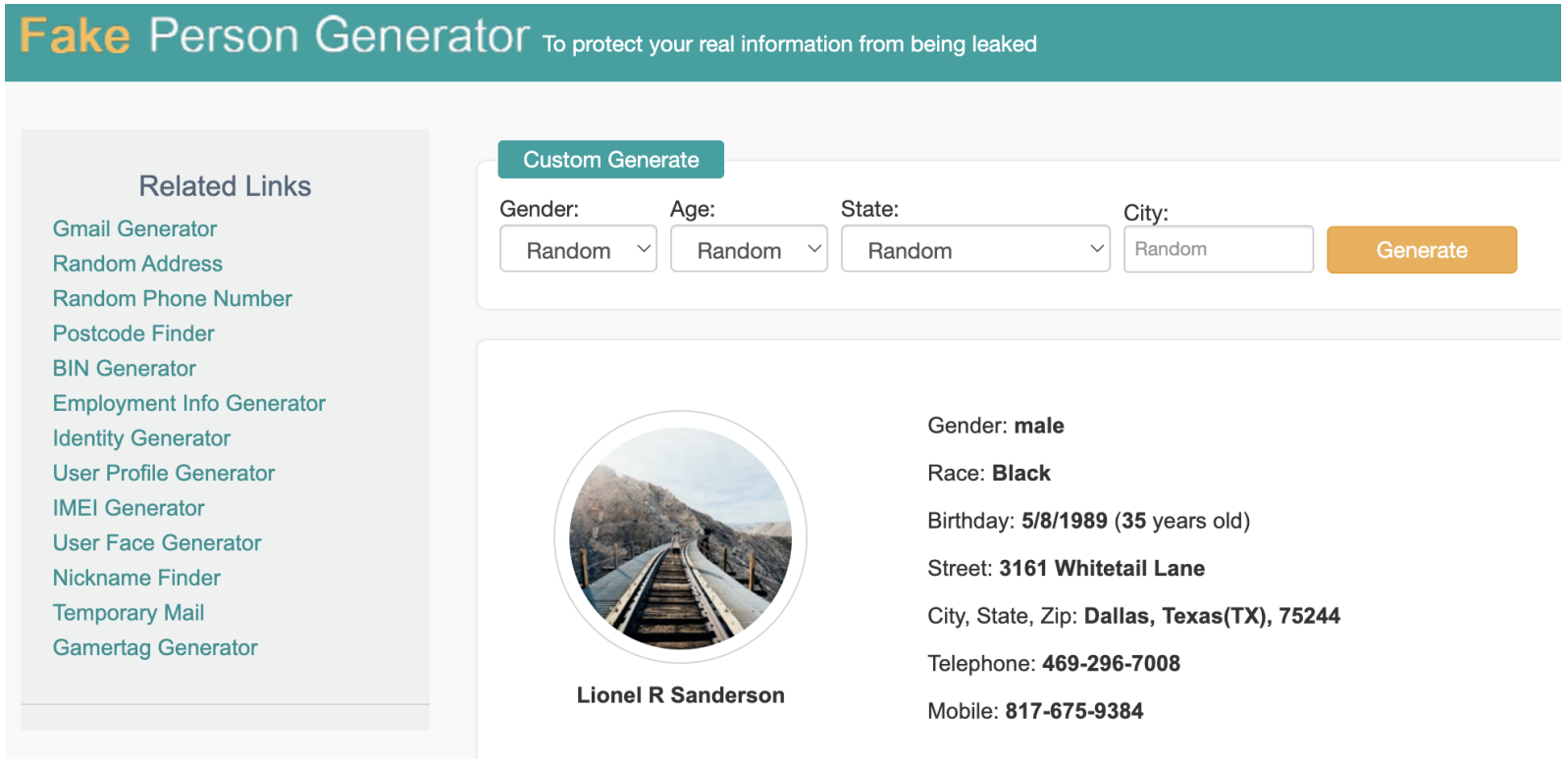


Synthetic Identies

Synthetic Identities – Sock Puppet Accounts



https://this-person-does-not-exist.com



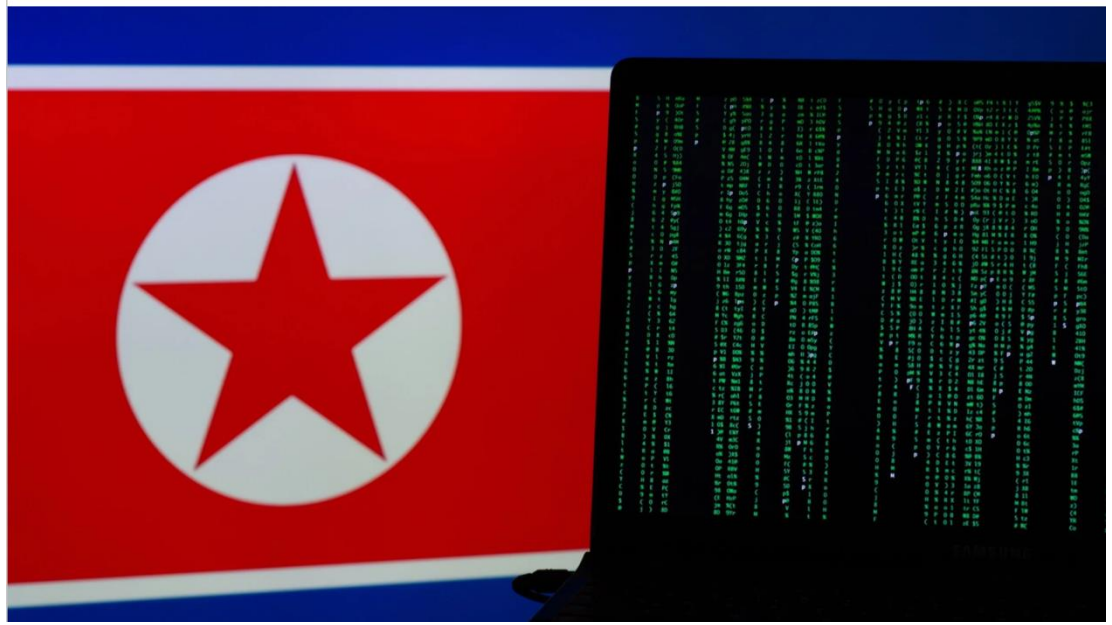
https://www.fakepersongenerator.com

Synthetic Identity

Security Firm Discovers Remote Worker Is Really a North Korean Hacker

Security awareness company KnowBe4 noticed something was fishy when the employee's company-issued Mac began loading malware.

Jul 23, 2024



- Fastest growing type of fraud
- >$1b in 2023
- Expected $40b by 2027
- Face swaps Injection attacks increased 704% in H2 vs H1 2023
- Yahoo Boys = Romance Scams
- NK hired employees in Tech orgs
- Now that’s an interesting story…

KnowBe4 & Synthetic Identities

X Post Button

23
Jul

How a North Korean Fake IT Worker Tried to Infiltrate Us

Stu Sjouwerman

Post Share Share 443

Incident Report Summary: Insider Threat

First of all: *No illegal access was gained, and no data was lost, compromised, or exfiltrated on any KnowBe4 systems. This is not a data breach notification, there was none. See it as an organizational learning moment I am sharing with you. If it can happen to us, it can happen to almost anyone. Don't let it happen to you.* We wrote an FAQ*, answering questions from customers. Story updated 7/27/2024.*

TLDR: KnowBe4 needed a software engineer for our internal IT AI team. We posted the job, received resumes, conducted interviews, performed background checks, verified references, and hired the person. We sent them their Mac workstation, and the moment it was received, it immediately started to load malware.

KnowBe4 & The North Korean Employee

- Applied for Principal Software Engineer with a fake identity, appeared as a US citizen of Asian descent.
- Claimed education in Hong Kong and work experience at well-known US companies.
- Passed interviews, technical, reference, and background checks, then hired.
- Received Apple laptop with FIDO-enabled Yubikey, requested laptop to be shipped to a new location.
- July 15, 2024: Attempted to install password-stealing malware but failed multiple times.
- Generated EDR alerts; KnowBe4 SOC contacted employee who made excuses and refused an audio session.
- Laptop isolated 25 minutes after first alert.
- Incident data shared with FBI, confirmed as a North Korean fake employee.
- Public blog post released; story went viral with extensive press coverage.
- Multiple companies shared similar experiences, leading to webinars, a whitepaper, and educational articles.

Stock Photo

AI-Photo
Blending of NK emp.

10 Things KnowBe4 Learned from the NK Hiring Event

1. Trained recruiters to spot DPRK resume red flags.
2. Provided recruiters phone carrier lookup tools.
3. Phone-based screening required and not just email
4. Recruiters search for applicant's public social media profiles.
5. Updating identity verification to government standard.
6. Video interviews require camera on without background filters.
7. Recruiters trained to ask casual, location-based questions.
8. CISO consults if suspicion arises during interviews.
9. Equipment shipped only to verified addresses or UPS stores.
10. Internet searches of addresses on suspicious resumes.

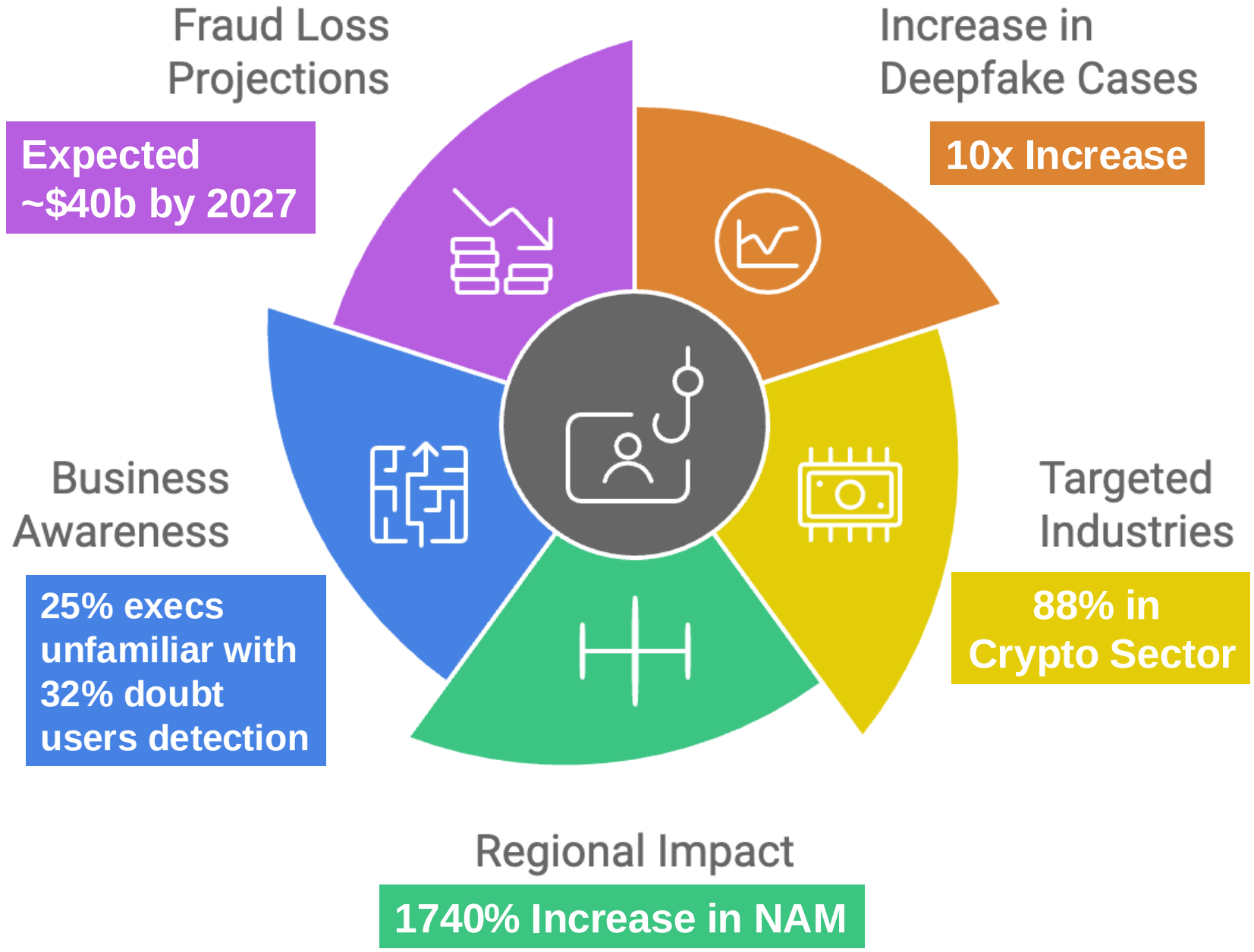
Source: https://blog.knowbe4.com/north-korean-it-worker-threat-10-critical-updates-to-your-hiring-process


An Algo-rithim

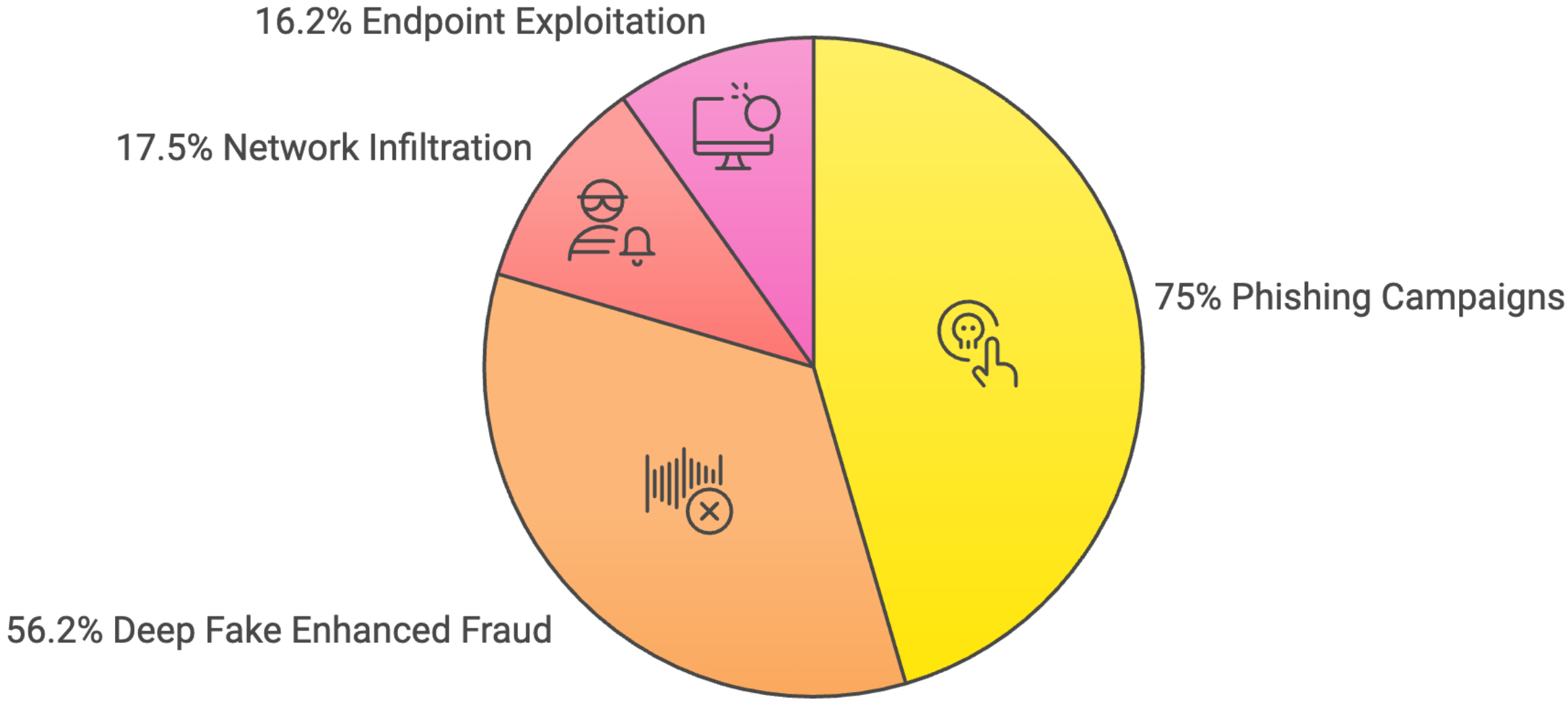

on Edge
Never relaxed?
Edge
Edge
Edge
Edge
Edge
Microsoft
Microsoft

Detection / Prevention

Deepfake Scams



CISO Survey – AI Threats – Team 8

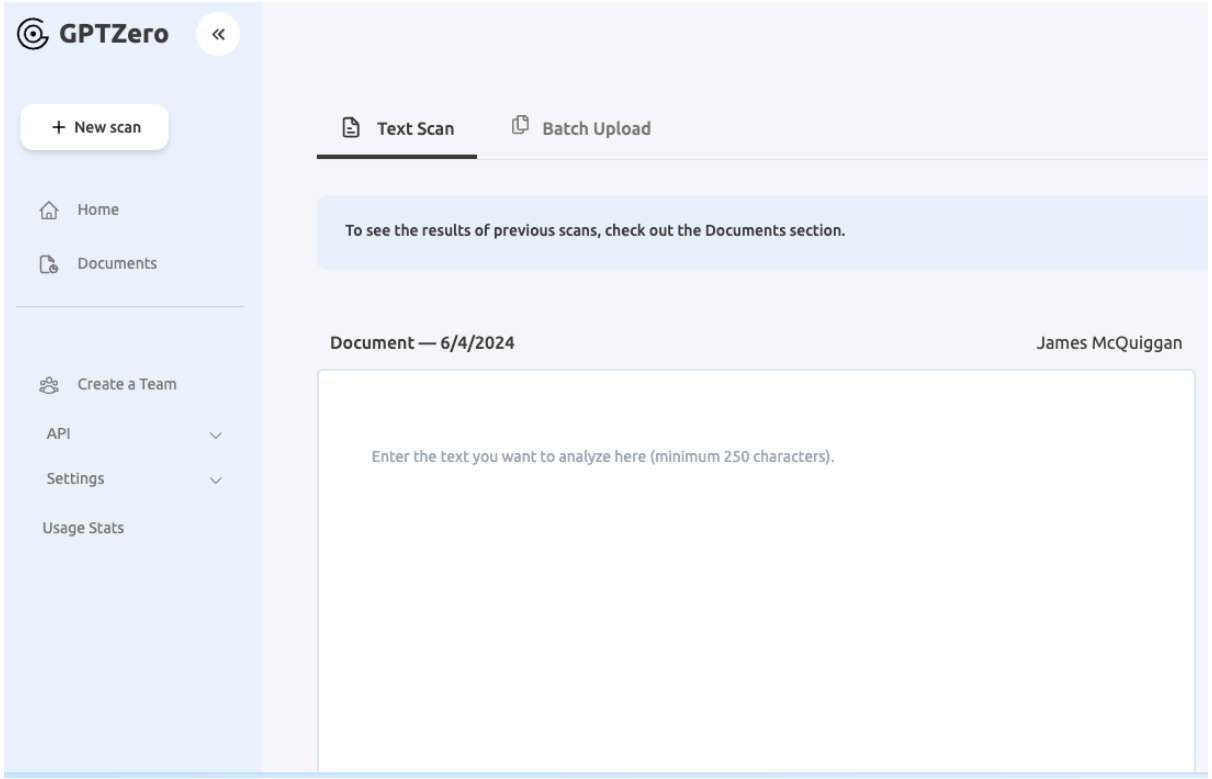


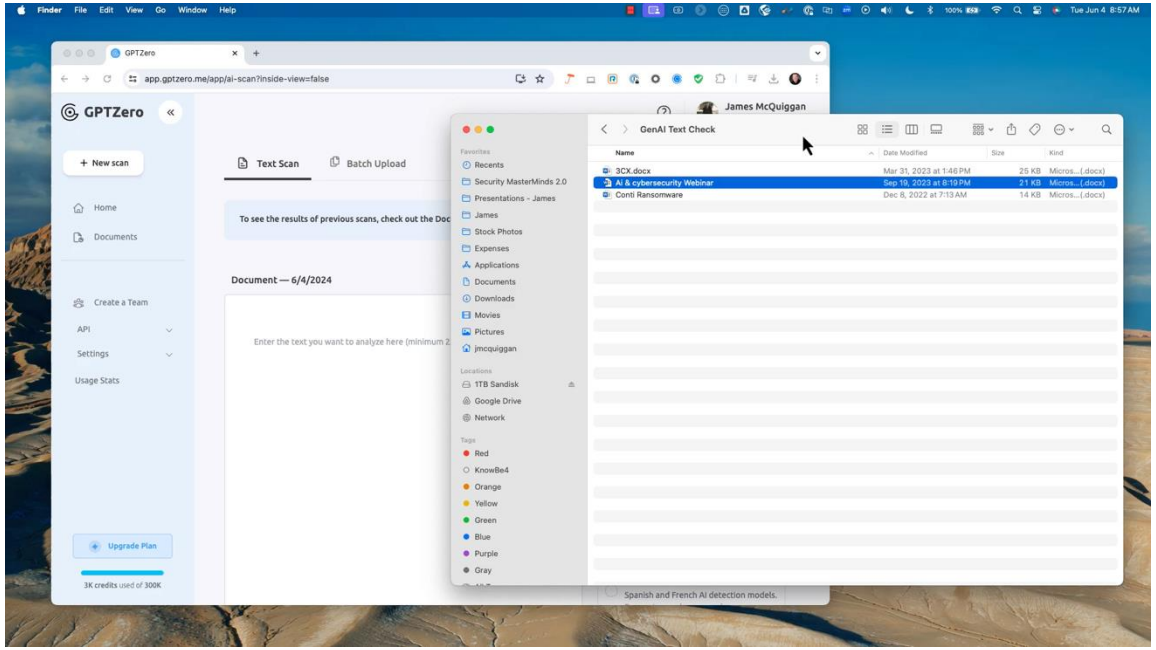
Basic Detection

- The criminals have leveled up
- We need to continue the same path… with some adjustments
- People, Processes, Technology
- People: If interacting, consider code words or ask unusual questions
- Technology: Detectors
- Processes: Frameworks

GPTZero – GenAI Synthetic Text Detection

- https://app.gptzero.me/app/ai-scan





Synthetic Video Detection Challenges

- Non-real time
- Not full-proof
- No standard detection method yet
- Generation tech advances outpace detection tech
- False Positives are plentiful
- Still requires manual labor

ST Engineering launches Einstein.AI deepfake detector

By Adam Campbell Last updated September 5, 2024

Deepfake Detector

deepware®

DEEPFAKE DASHBOARD™

Source: https://arxiv.org/pdf/2301.05819.pdf

deepware®

API

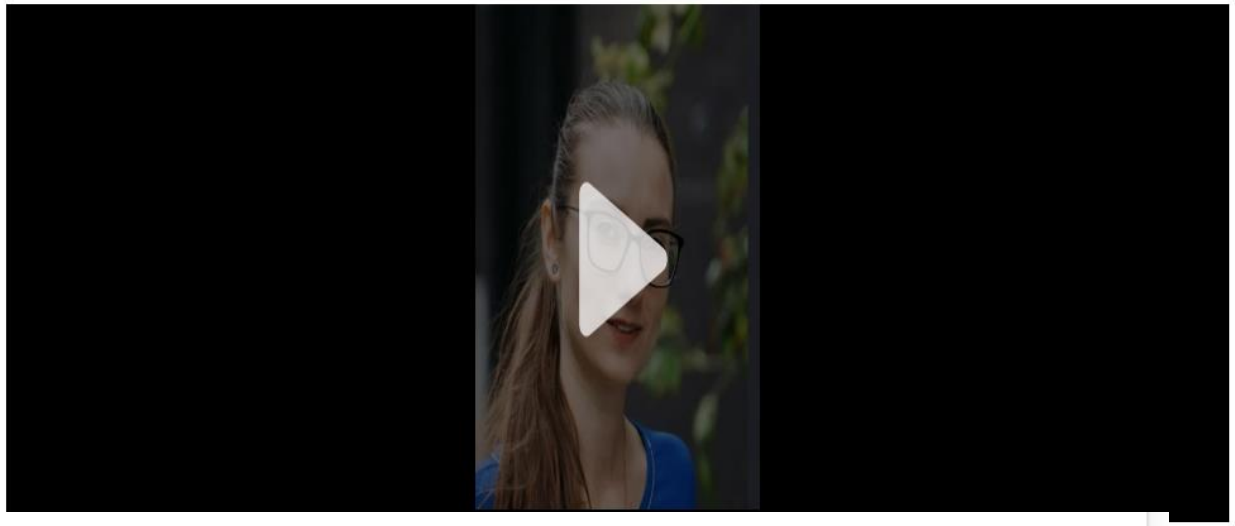
GitHub

NO DEEPFAKE DETECTED

Name: af66f232-b4ac-4a35-a2dd-d6c93655ba68.mp

Size: 9.5 MB

Deepware aims to give an opinion about the scanned video and is not responsible for the result. As Deepware Scanner is still in beta, the results should not be treated as an absolute truth or evidence.



Model Results

Avatarify: NO DEEPFAKE DETECTED(43%)

Deepware: NO DEEPFAKE DETECTED(0%)

Seferbekov: NO DEEPFAKE DETECTED(1%)

Ensemble: NO DEEPFAKE DETECTED(0%)

Video

Duration: 37 sec

Resolution: 512 x 512

Frame Rate: 30 fps

Codec: h264

Audio

Duration: 37 sec

Channel: mono

Sample Rate: 48 khz

Codec: aac

James HeyGen Video - Deepware

Deepware aims to give an opinion about the scanned video and is not responsible for the result. As Deepware Scanner is still in beta, the results should not be treated as an absolute truth or evidence.

DEEPFAKE DETECTED

Name: James -BSidesCPH.mp4

Size: 5.5 MB

User

Source

202

Model Results

Avatarify: DEEPFAKE DETECTED(94%)

Deepware: NO DEEPFAKE DETECTED(0%)

Seferbekov: NO DEEPFAKE DETECTED(31%)

Ensemble: NO DEEPFAKE DETECTED(4%)

Video

Duration: 9 sec

Resolution: 1920 x 1080

Frame Rate: 25 fps

Codec: h264

Audio

Duration: 9 sec

Channel: stereo

Sample Rate: 48 khz

Codec: aac

o

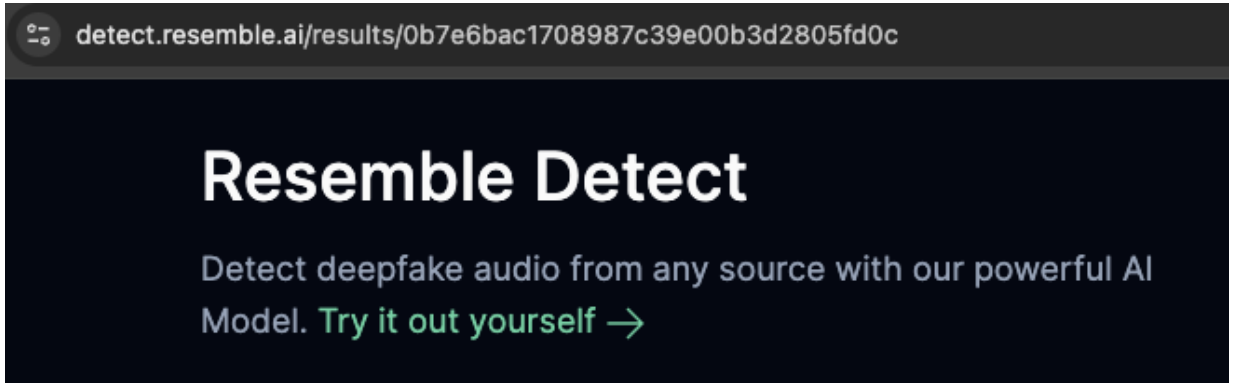
n: 9 sec

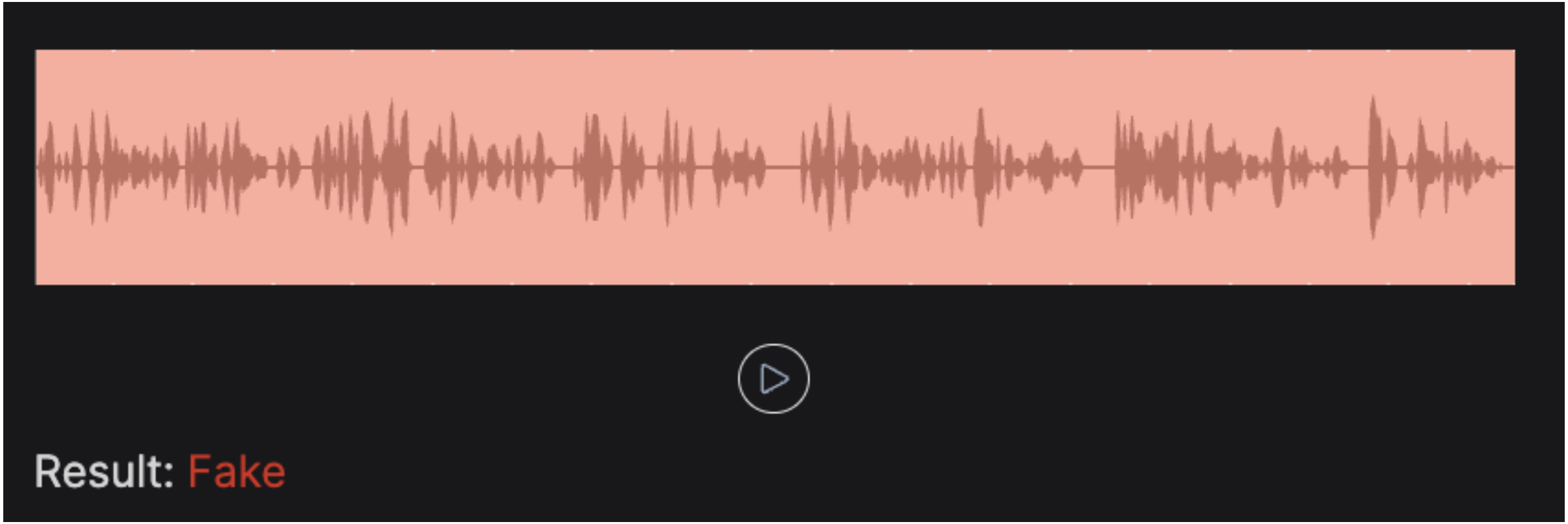
l: stereo

Rate: 48 khz

aac

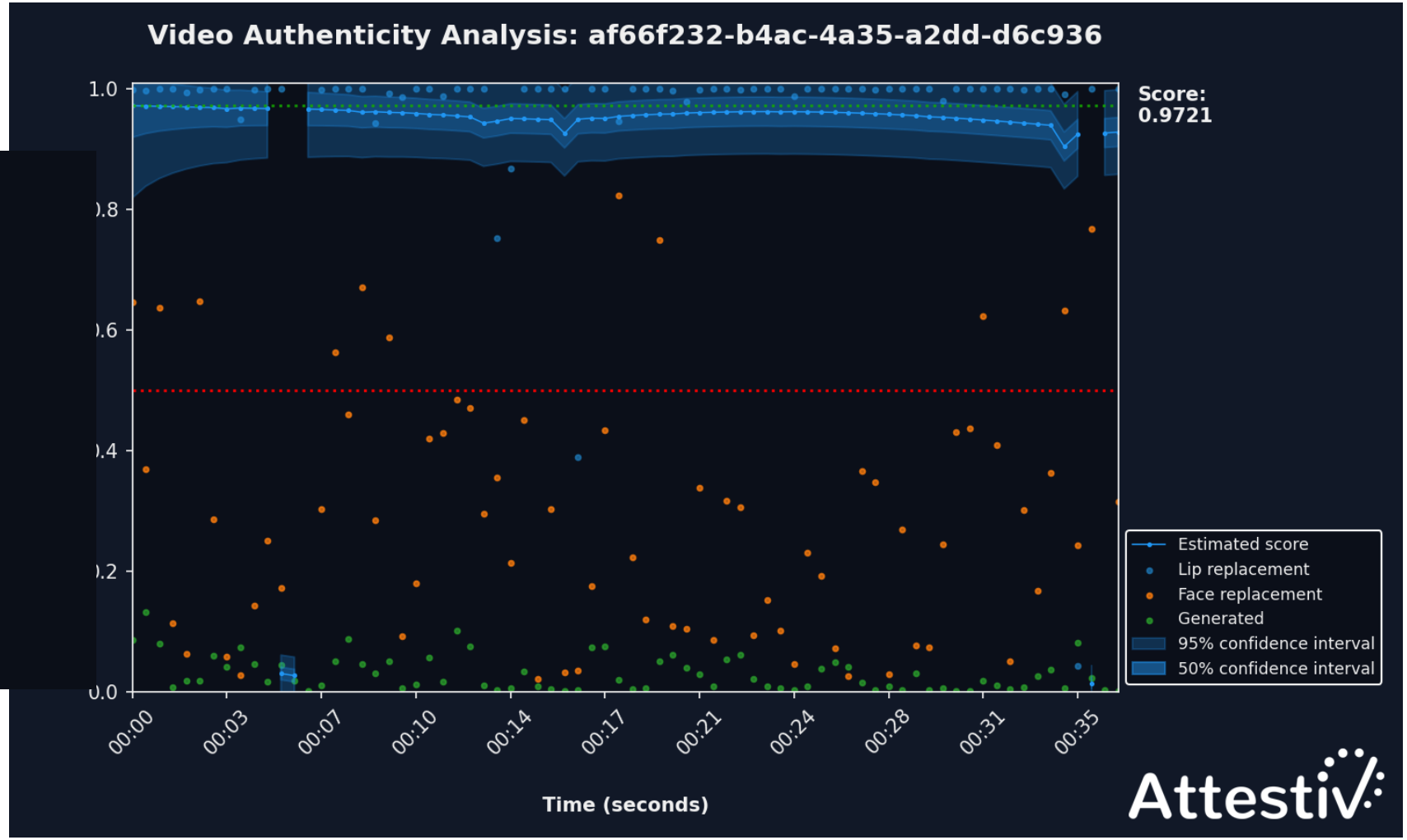
Andrada’s Audio File





https://detect.resemble.ai/results/0b7e6bac1708987c39e00b3d2805fd0c

Attestiv



Overall Suspicion Rating 97

Our analysis of the video shows that the overall suspicion rating is 97, which means that parts of the video are **VERY SUSPICIOUS***

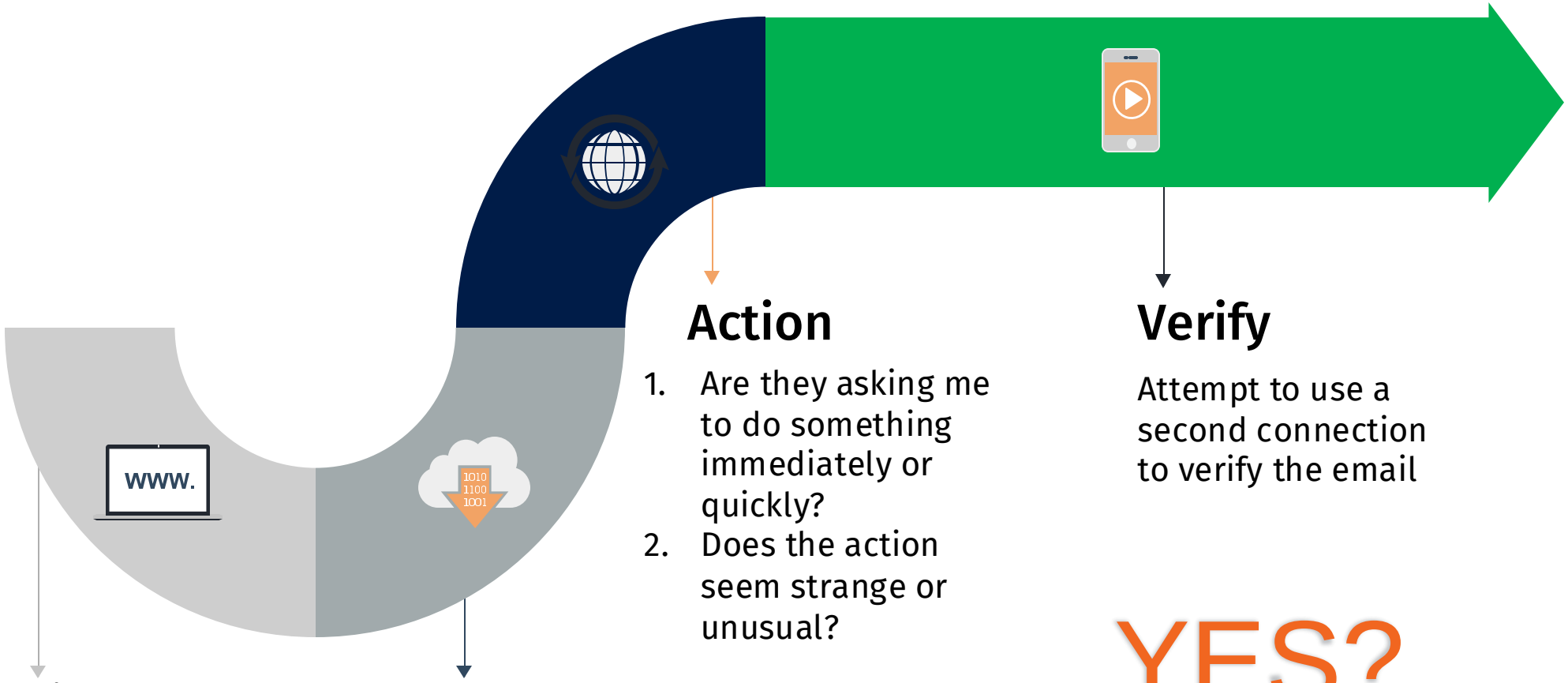
* Please note that the Overall Suspicion Rating (OSR) is an estimate based upon our analysis, and not a guaranteed or definitive statement of the presence of editing or deepfakes. Please click here to learn more about the OSR and Attestiv's stance on responsible AI. You can also learn more about our analysis and interpreting our results by visiting our FAQ page here.

10 Best AI DeepFake Detector Tools

Additional Deepfake Detectors

- Sentinel.ai – Requires a demo
- Sensity.ai – Requires Corporate email and detailed reason to use their platform
- Oz Forensics – Facial recognition
- DuckDuckGoose – demo required
- Deepware – free tool – False Postive
- Attestiv.com – free tool – 65%

Process: 3 Questions for Email / Synthetics



Email

Is the email unexpected?

Sender

Is this person a stranger? #StrangerDanger

Action

1. Are they asking me to do something immediately or quickly?
2. Does the action seem strange or unusual?

Verify

Attempt to use a second connection to verify the email

YES?

Process: Synthetic Video Tips (VeSSPER)

Verify	• Ask questions or get them to do something unpredictable like writing a specific word on paper and showing it on camera.
Skepticism	• Be cautious if someone you've only met online requests money, personal information or any other sensitive details.
Secure	• Use secure, encrypted apps for texting and voice
Privacy	• Protect personal information available publicly
Education	• Keep up to date with newsletters, podcasts etc.
Report	• Report it to the relevant authorities like ic3.gov or police

"Don't ask, 'Is this real?' Ask, 'Why does this exist?'" - FAIK

What Should We Be Asking?

✗ Is this a deepfake?

✓ Consider these questions…

What Should We Be Asking?

- Why does this exist?
- What story is it telling?
- Who Benefits?
- What are the possible goals?

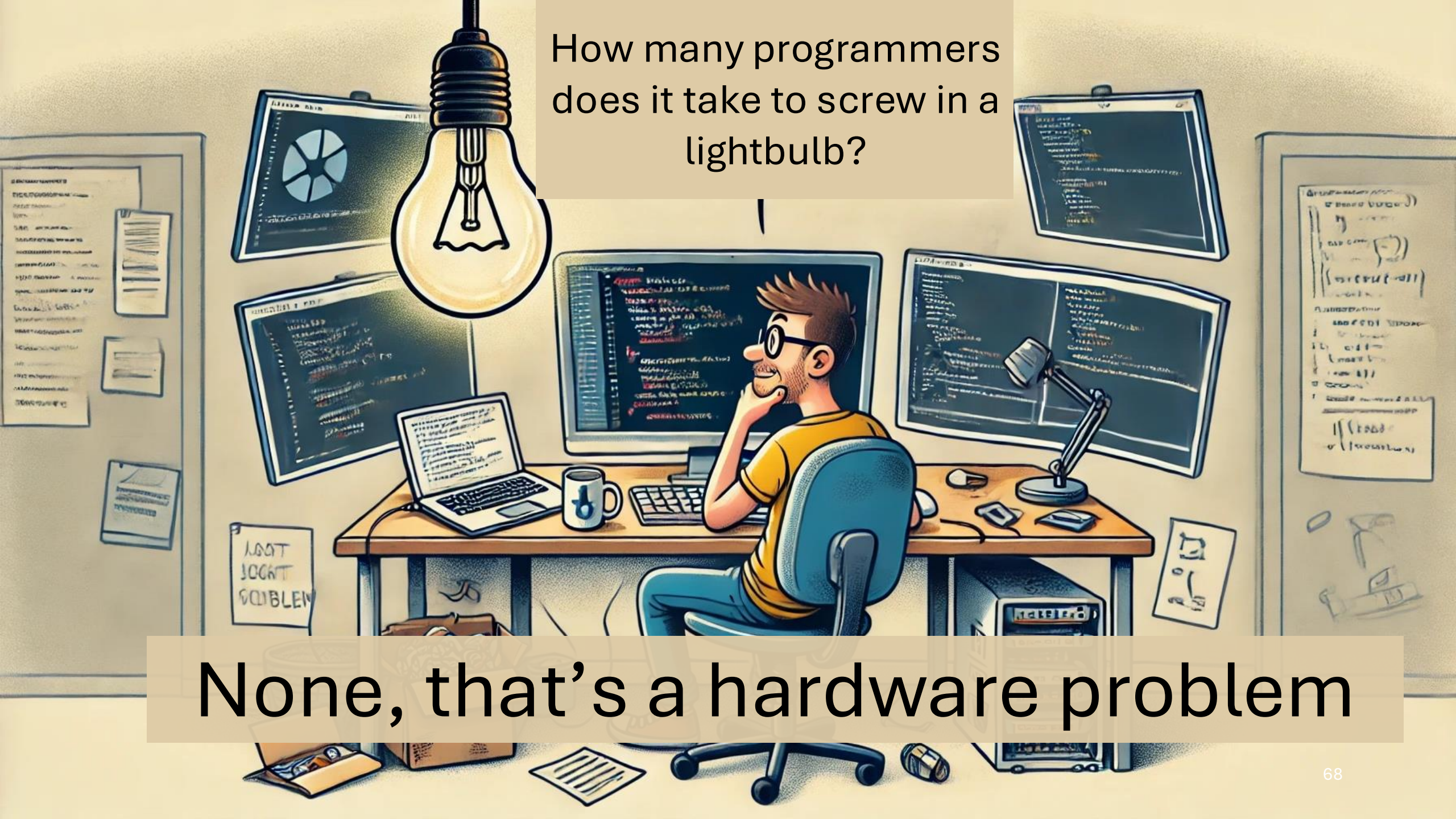
Apply the FAIK Factor Framework

F Freeze & Feel

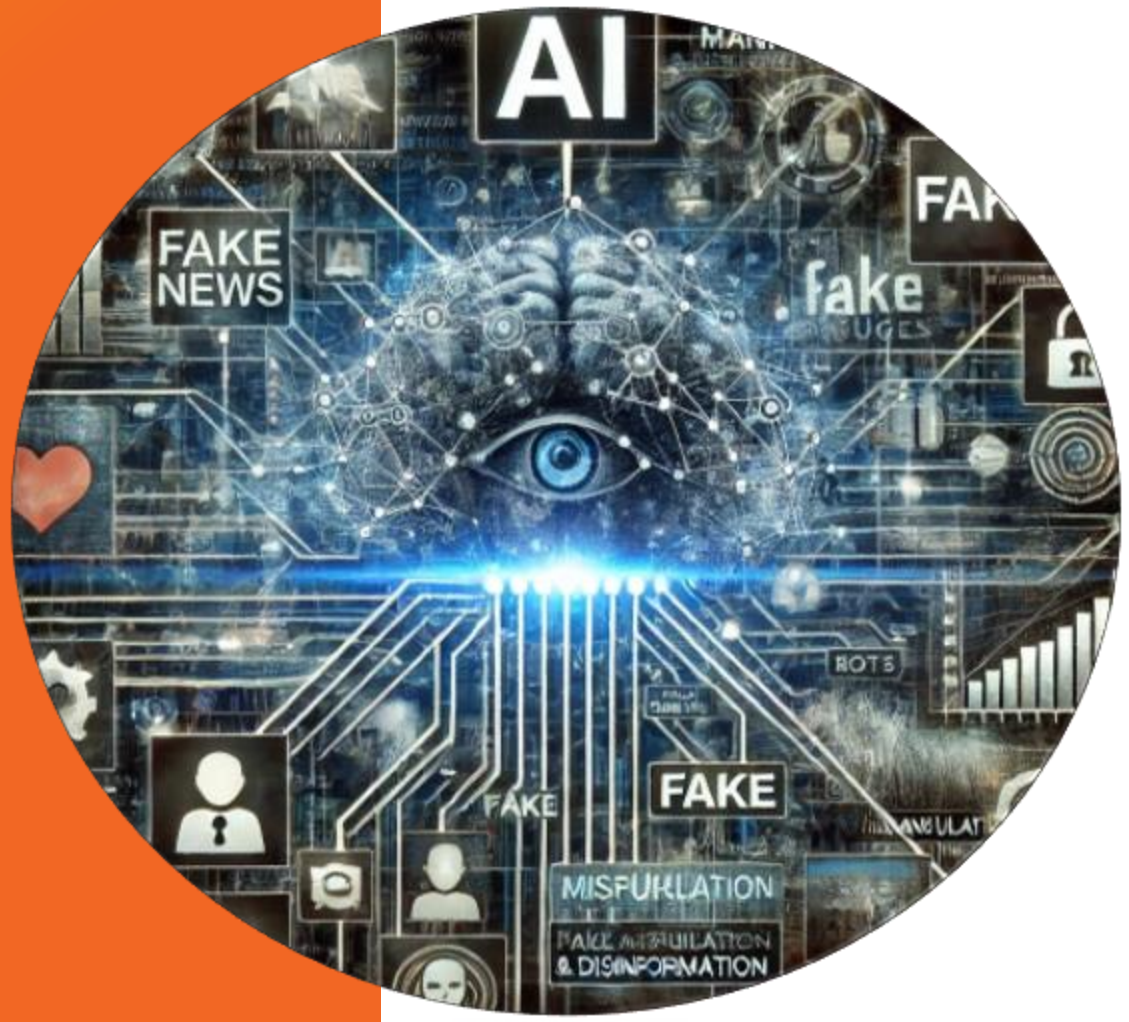
A Analyze the Narrative & Emotional Triggers

I Investigate (claims, sources, etc.)

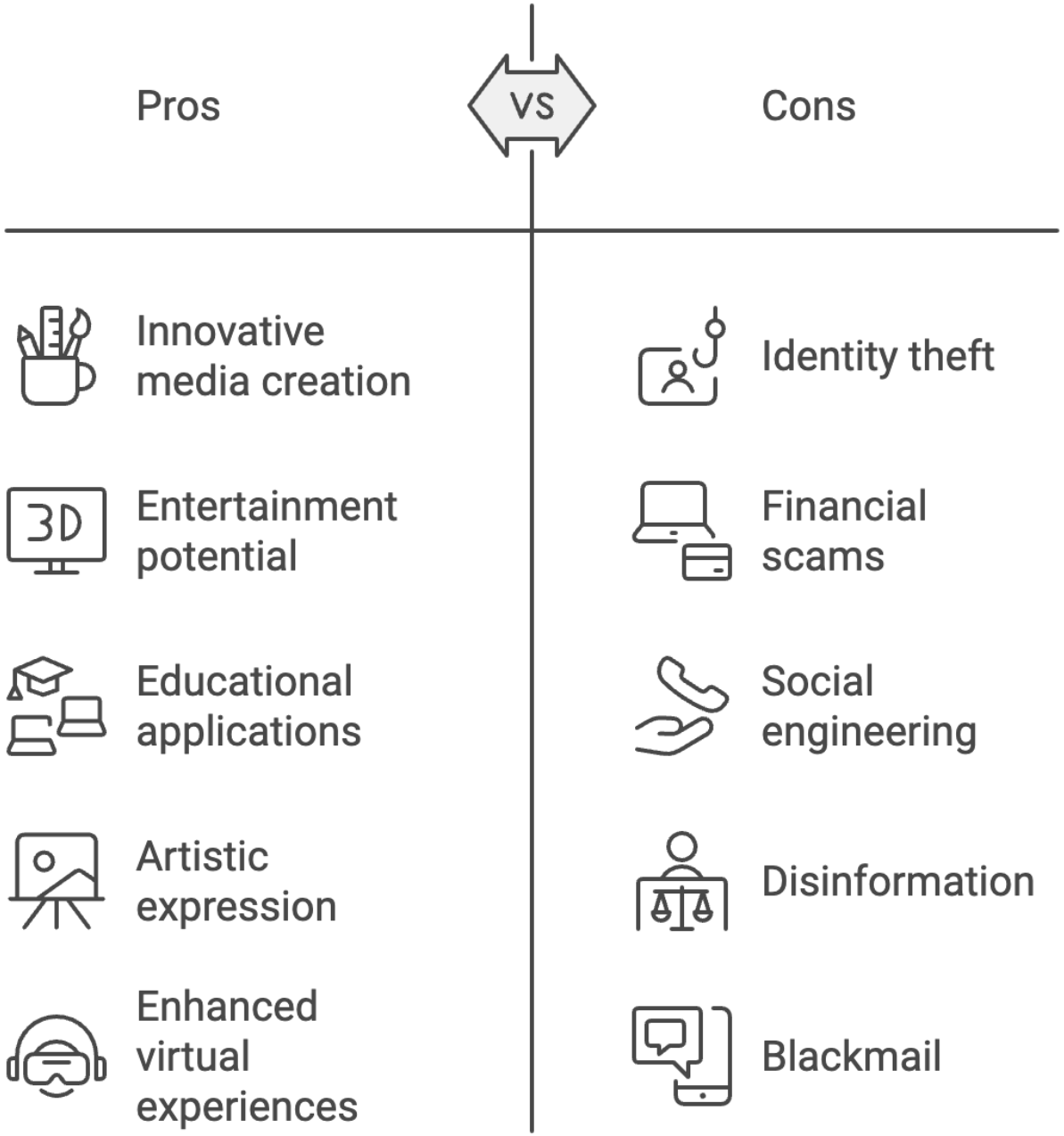
K Know, confirm, and keep vigilant


How many programmers does it take to screw in a lightbulb?
None, that’s a hardware problem

Final Thoughts / Wrap-up



Deepfakes Duality



Takeaways

AI is an incredible tool available to all – Ensure we're educating everyone. **Politely Paranoid**

Be aware of AI Hallucinations, Biases and Deepfakes

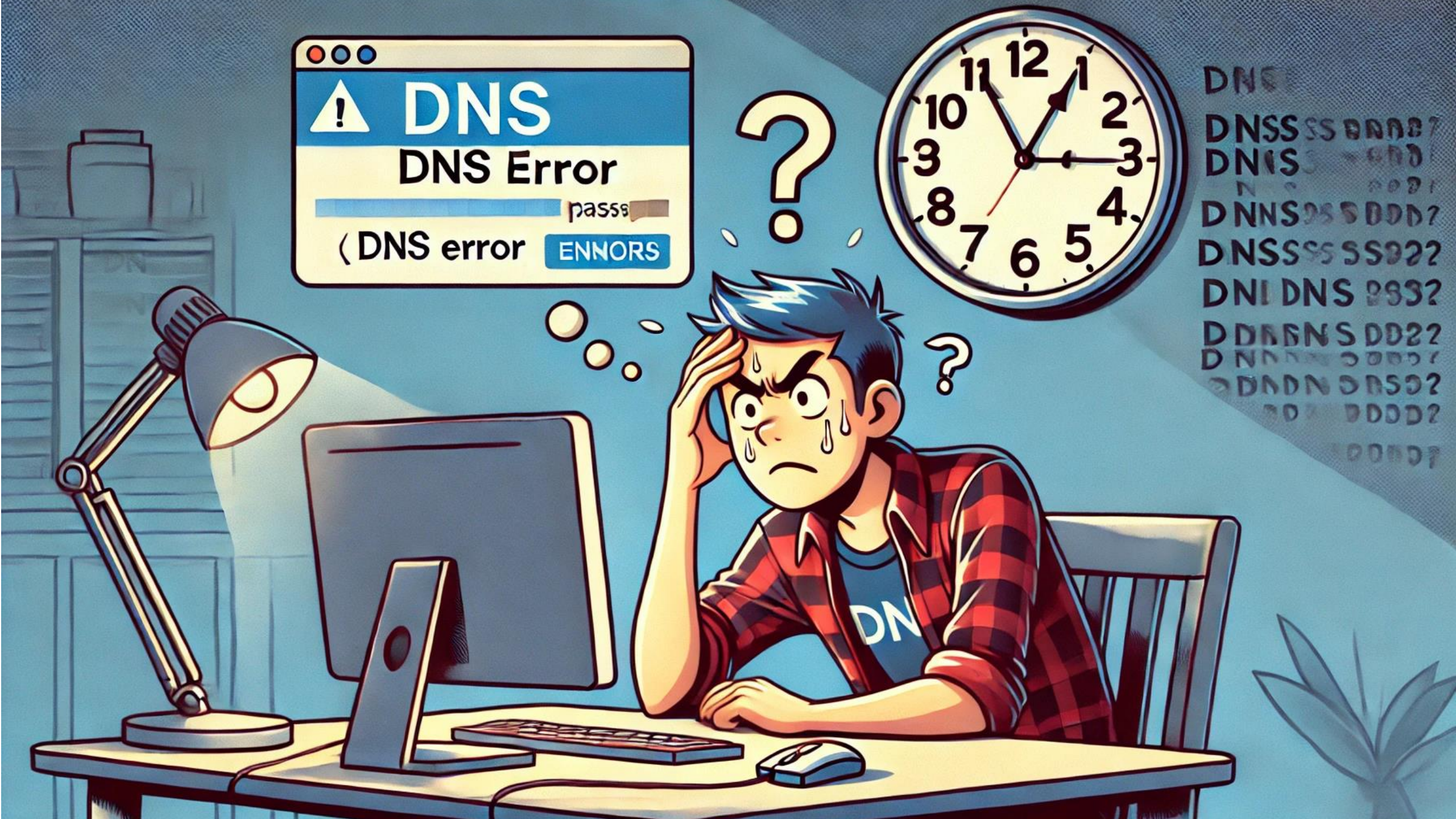
Trust AND Verify

The Phishing game hasn't changed. Be aware, don't rush, check links, be **skeptical**

Final Thoughts

Healthy Skepticism

Politely Paranoid


DNS
DNS Error
pass
DNS error
ENNORS

Thank You For Your Attention

James R. McQuiggan, CISSP, SACP

Email: jmcquiggan@knowbe4.com

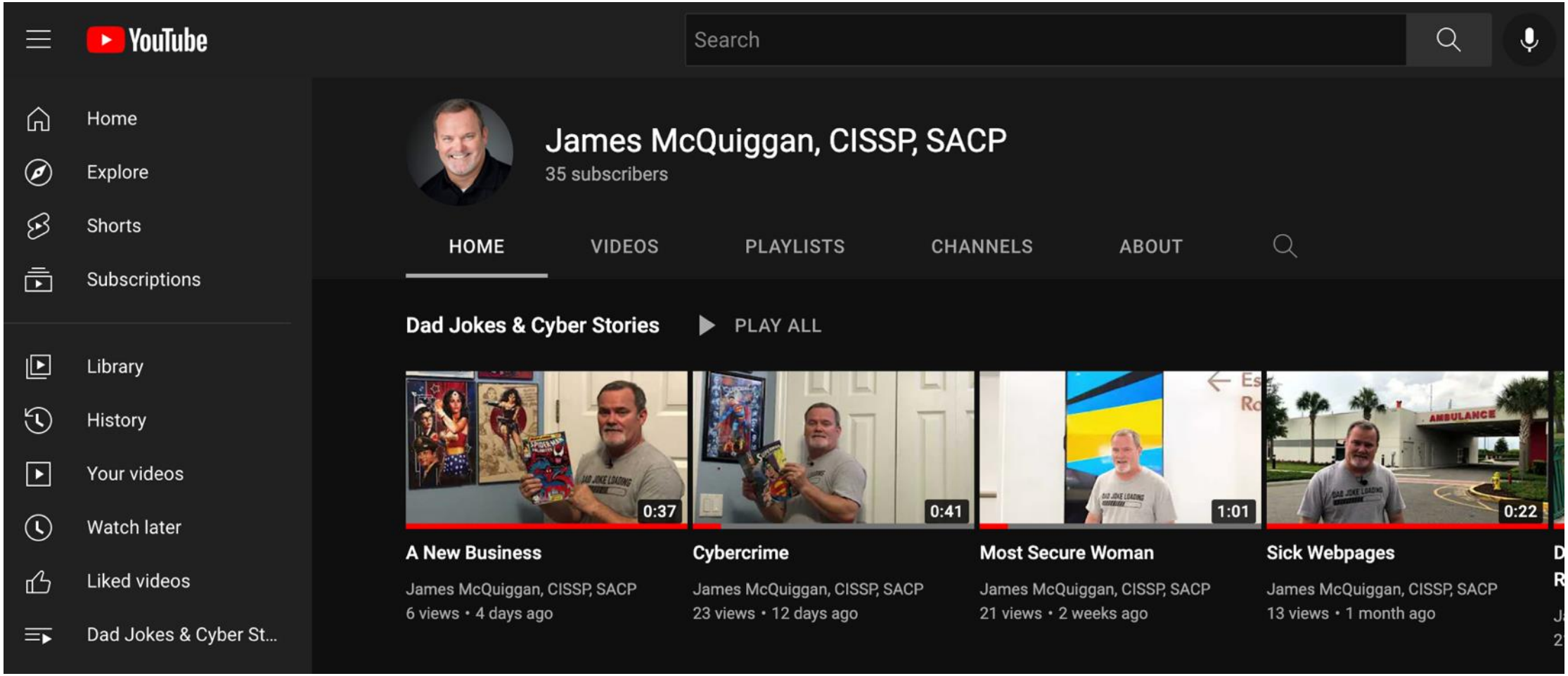
KnowBe4 Blog: blog.knowbe4.com

Connect with Me!

LinkedIn jmcquiggan

X @james_mcquiggan

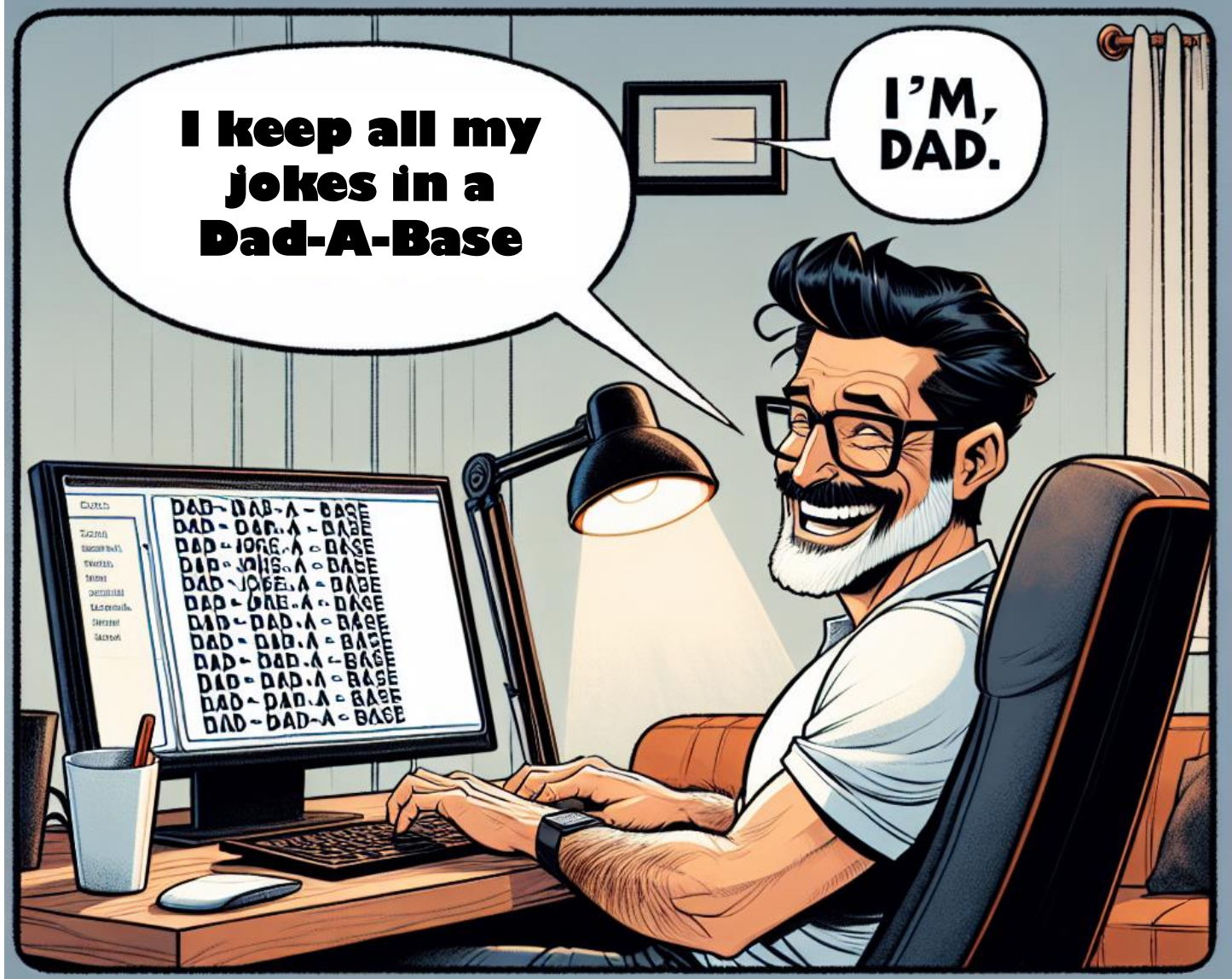
Website jamesmcquiggan.com



YouTube: James McQuiggan and Dad Jokes

https://www.youtube.com/@JamesMcQuigganCISSP

Yes… I have a way of keeping track of my **Dad Jokes**



Resources

- Resemble Audio - https://detect.resemble.ai/
- Deepware Detector - https://scanner.deepware.ai/
- HeyGen Video - https://app.heygen.com
- Hedra - https://www.hedra.com
- 11Labs – https://elevenlabs.io/
- PlayHT - https://play.ht/

- **Presentation**: https://github.com/jmcquiggan/presentations

KnowBe4

THANK YOU!