

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. This presentation, and the following written materials, contain KnowBe4's proprietary and confidential information and is not to be published, duplicated, or distributed to any third party without KnowBe4's prior written consent. Certain information in this presentation may contain "forward-looking statements" under applicable securities laws. Such statements in this presentation often contain words such as "expect," "anticipate," "intend," "plan," "believe," "will," "estimate," "forecast," "target," or "range" and are merely speculative. Attendees are cautioned not to place undue reliance on such forward-looking statements to reach conclusions or make any investment decisions. Information in this presentation speaks only as of the date that it was prepared and may become incomplete or out of date; KnowBe4 makes no commitment to update such information. This presentation is for educational purposes only and should not be relied upon for any other use.

Product and Tool References Disclaimer: Any products, tools, software, or services referenced, demonstrated, or mentioned in this presentation are provided for informational and educational purposes only. The inclusion of such references does not constitute an endorsement, recommendation, or approval by KnowBe4 or the presenter. KnowBe4 and the presenter have not received any compensation, consideration, or other financial benefit from the referenced product or service providers. Attendees should conduct their own independent evaluation and due diligence before using any referenced products or services. KnowBe4 and the presenter disclaim any responsibility or liability for the performance, security, or suitability of any referenced products or services.

DISCLAIMER



Is it Live or is it Memorex?
Memorex is a deepfake!



@31337magician



dylan-baklor



THE MAGICIAN





How long do you
think it took me to
make these videos?

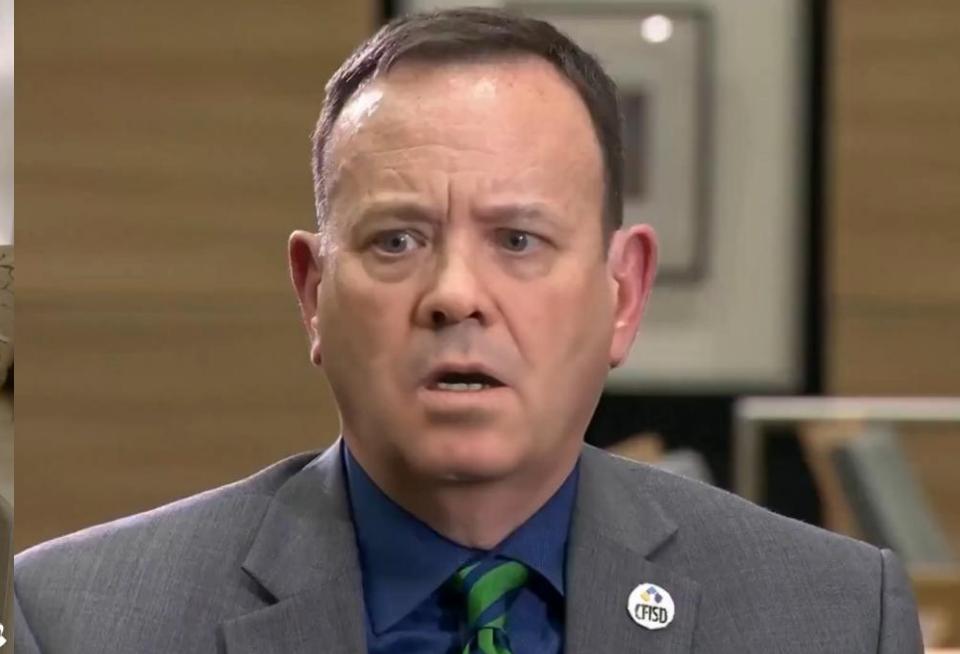
James R. McQuiggan

Professor, CTI, Full Sail University
Host, Simply Secured Podcast



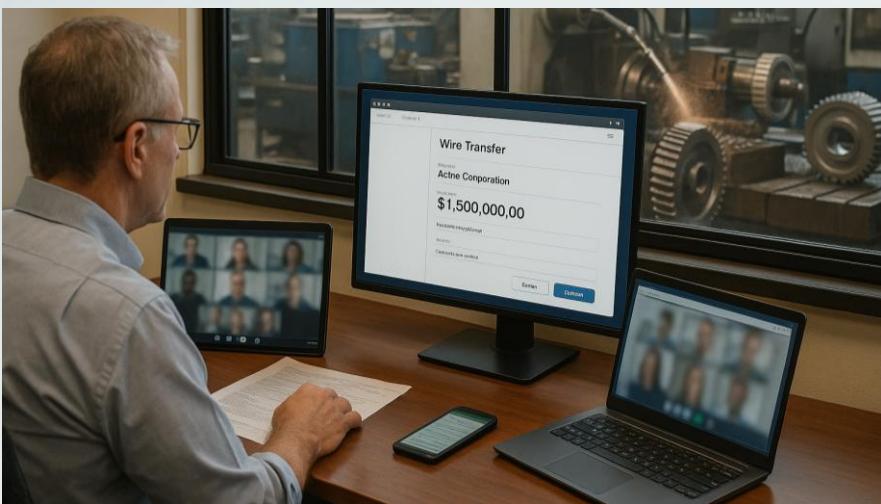


When I started messing with Deepfakes



Why Deepfakes Matter in Cybersecurity

The screenshot shows a news article from The Financial Times. The header reads "FINANCIAL TIMES" with a sub-header "COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI". Below the header, there's a "Cyber Security" section with a "Add to myFT" button. The main headline is "Arup lost \$25mn in Hong Kong deepfake video conference scam". A sub-headline states "UK-based engineering group identified as target of fraud that used digitally cloned CFO to trick staff".



Financial Fraud

Involves scams and wire transfer manipulation



Operational Disruption

Includes supply chain and IT support issues



Reputation Damage

Covers fake press releases and crisis statements



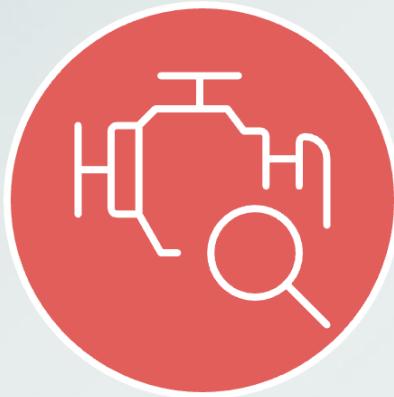
Regulatory Risk

Encompasses disclosure failures and compliance issues



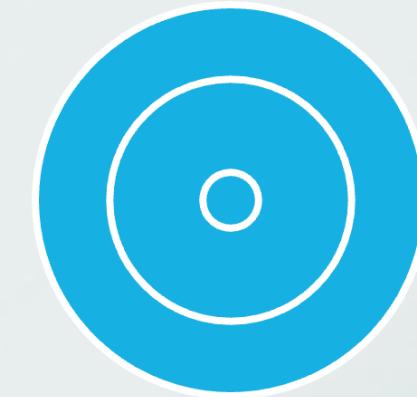
A Quick Look at Synthetic Media

Which network excels?



Generator

Creates synthetic data



Discriminator

Evaluates data authenticity

Synthetic Text - Malicious LLMs



Phishing LLM – Nation state Activity - LAMEHUG

The Hacker News

Home Data Breaches Cyber Attacks Vulnerabilities Webinars Expert Insights Contact



CERT-UA Discovers LAMEHUG Malware Linked to APT28, Using LLM for Phishing Campaign

Jul 18, 2025 Ravie Lakshmanan

Cyber Attack / Malware

Attribution:

- Russian APT28 (Fancy Bear/Forest Blizzard)
- ZIP archives containing three LAMEHUG variants
- Hugging Face: Qwen2.5-Coder-32B-Instruct

Attack Vector

- Phishing emails from compromised (BEC) government account
- Impersonating ministry officials
- Targeting executive government authorities

Technical Details

- Leverages Alibaba Cloud's coding-focused LLM
- Available on Hugging Face and Llama platforms
- Malicious command generation with python

Timeline

- First reported to CERT-UA on July 10, 2025
- This is an active, current threat

Recent LLM Attacks

Agentic AI coding assistant helped attacker breach, extort 17 distinct organizations



Cyber Press

World's First AI-Powered Ransomware 'PromptLock' Emerges Using GPT-OSS-20B

BY ANUPRIYA / AUGUST 28, 2025

Categories: Cyber Security News • Cybersecurity • Ransomware

```
lma/v1/chat/completions HTTP/1.1
2.42.0.253:8443
User-Agent: Go-http-client/1.1
Content-Length: 1665
Type: application/json
Session-Key: env_windows_2025-08-26-02-54-56;task_name=probe;
Content-Encoding: gzip

{"gpt-oss:20b","messages":[{"role":"system","content":"You are a Lua code generator. Generate Lua code wrapped in \u003ccode\u003e \u003ccode\u003e tags without any comments."},{"role":"user","content":"exfiltrate"}]}
```

**World's First AI-Powered Ransomware 'PromptLock'
Emerges Using GPT-OSS-20B**

The image shows a news article from Cyber Press about the emergence of the world's first AI-powered ransomware, 'PromptLock', which uses GPT-OSS-20B. The article includes a snippet of the generated Lua code and the headline.

Password Cracking – Latest Results From KnowBe4

Forthehopeofitall13\$

khamzatchimaev1!

Houseofdragon2894!

CastleArsenal123

LionsSuperbowl2223

Ofundamental@123A

Megan1009@@@@@@@

#LydiaLynne2019!

Soccerrugby110\$\$

KasselGermany1924

RaidShadowLegendsRocks

TheDreamJob4242!

Buddycannoli888!

Myfriendsaremypower323!!



Kevin Mitnick
@kevinmitnick



This is my new bad ass password cracker.

I have 24 4090's + 6 2080's all clustered running Hashtopolis.

Thanks to the awesome team at [@KnowBe4](#) that set up and configured the servers for me.

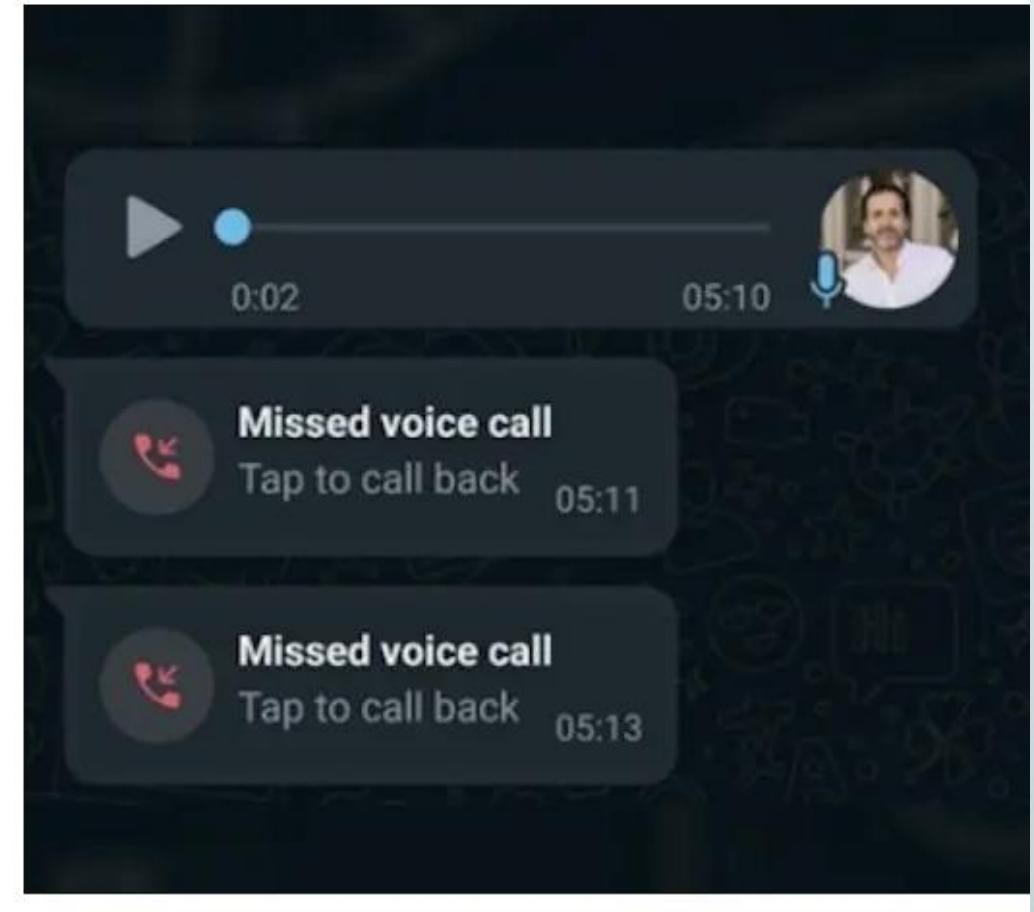
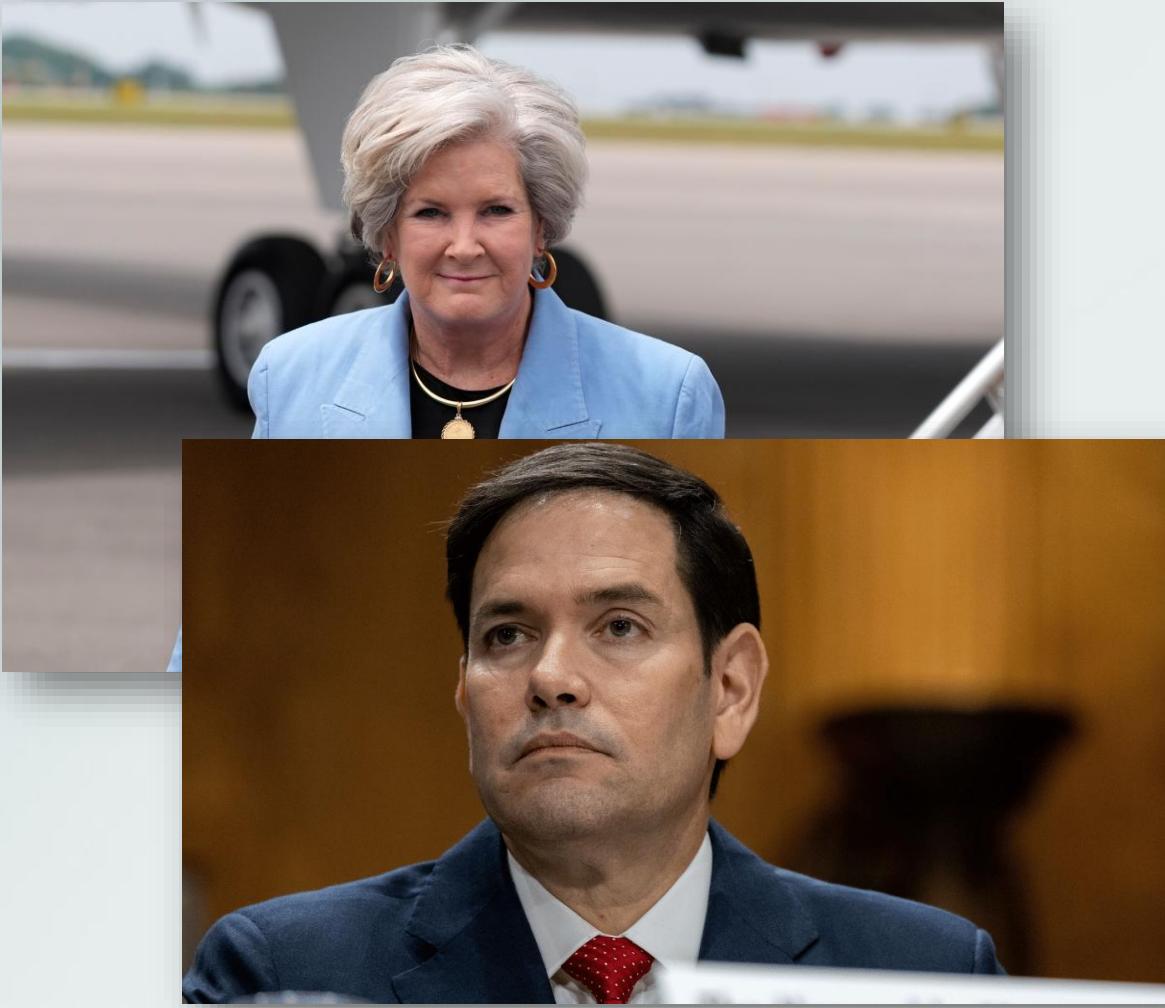
Now to go crack some hashes :-))))))



9:34 AM · Apr 21, 2023 · 458.1K Views



Synthetic Audio and Social Engineering

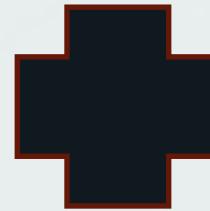


Kyle Wilhoit of Palo Alto Networks' Unit 42 division

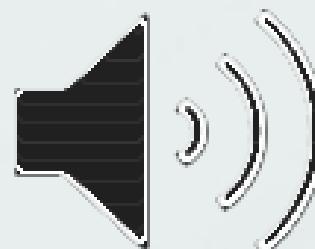
ChatGPT + Syn. Audio + Call Center = AI Social Engineering



OpenAI's New
ChatGPT



Call Center
Support
Software



Using PlayHT



Dr. Gerald Auger
Simply Cyber
(and friend)

Disinformation Campaigns



Hany Farid · Following
UC Berkeley Professor & GetReal Co-founder
14h · Edited ·

In just the last 12 hours, we at **GetReal** have been seeing a slew of fake videos surrounding the recent conflict between Israel and Iran. We have been able to link each of these visually compelling videos to Veo 3.

It is no surprise that as generative-AI tools continue to improve in photo-realism, they are being misused to spread misinformation and sow confusion.

One simple tip-off (for now at least) is that all of these videos are either exactly eight seconds in length or composed of short (eights seconds or less) clips composited together. Why eight seconds? This is the current maximum length that Veo 3 can generate a continuous shot. Other models have slightly longer limits but 8-10 seconds is typical.

This eight-second limit obviously doesn't prove a video is fake but should be a good reason to give you pause and fact-check before you re-share.



You and 362 others

21 comments · 51 reposts

The National Guard's post

The National Guard
5d ·

Several videos from "Bob" have been making the rounds online. They are fake. Red flags indicating a fake: The uniform name tape reads "Bob," his rank reads "E-6," and there is gibberish where "U.S. Army" should appear. In one video, the AI-generated "man" even eats a burrito through a mask.

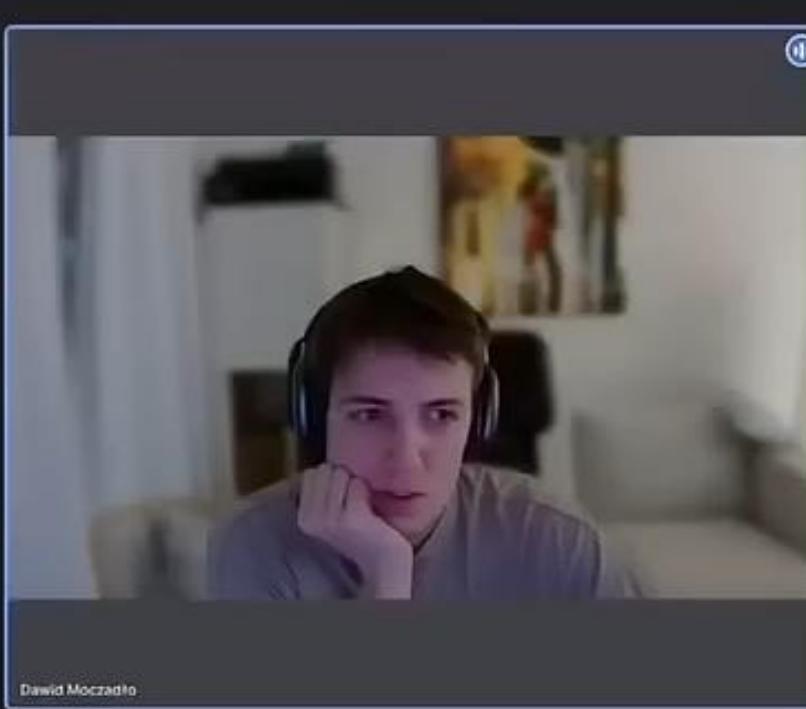
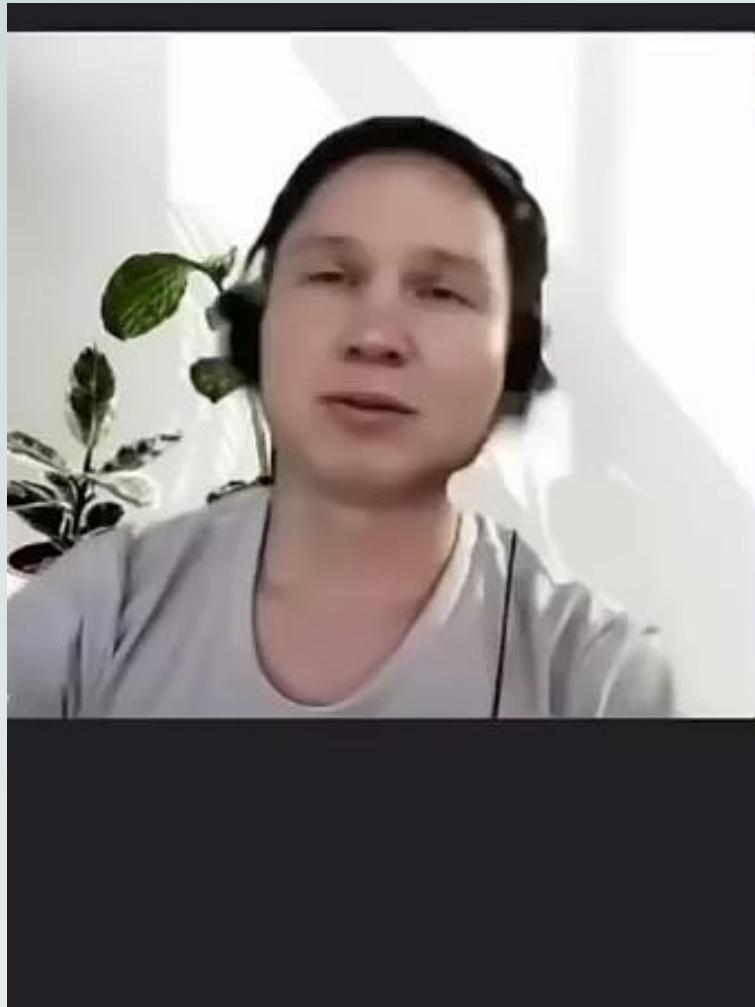
<https://www.nationalguard.mil/.../guard-identifies-ai...>

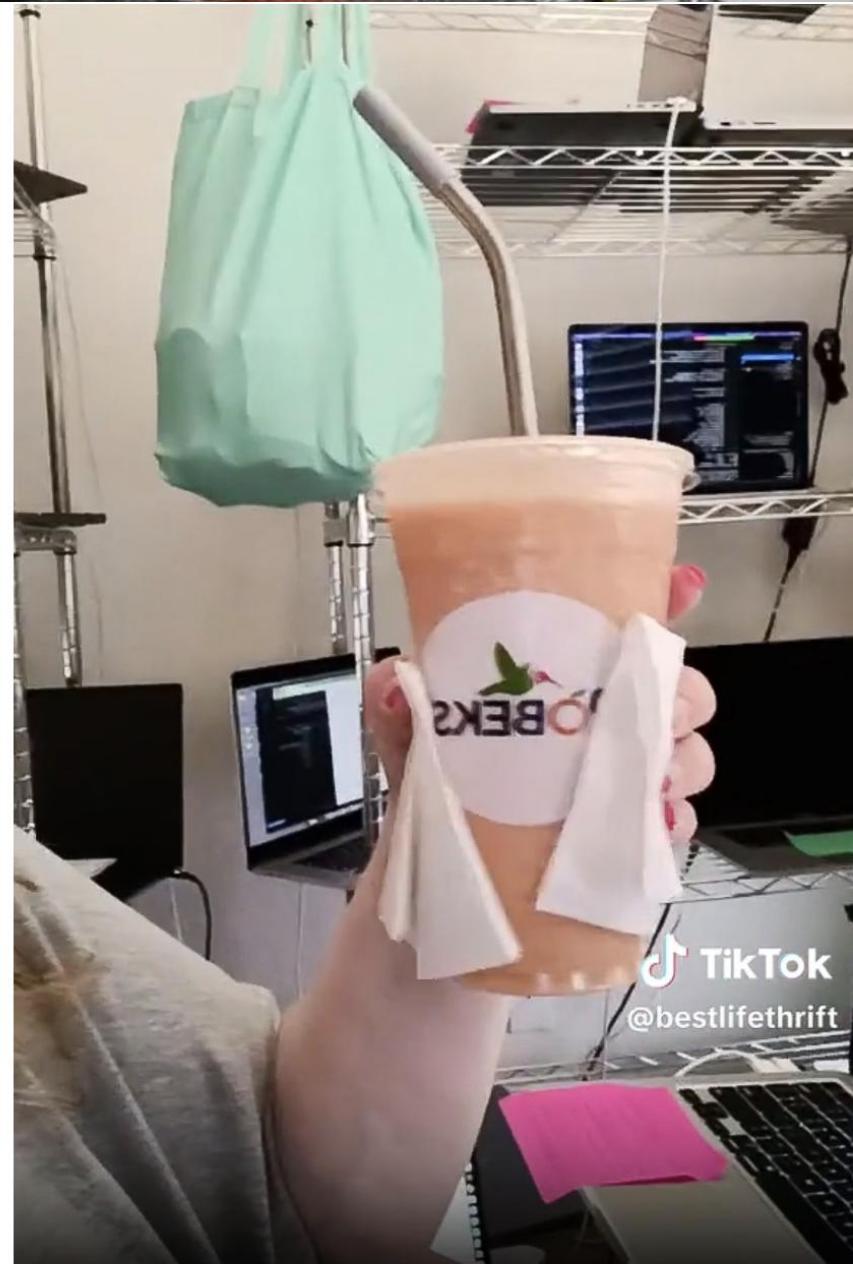
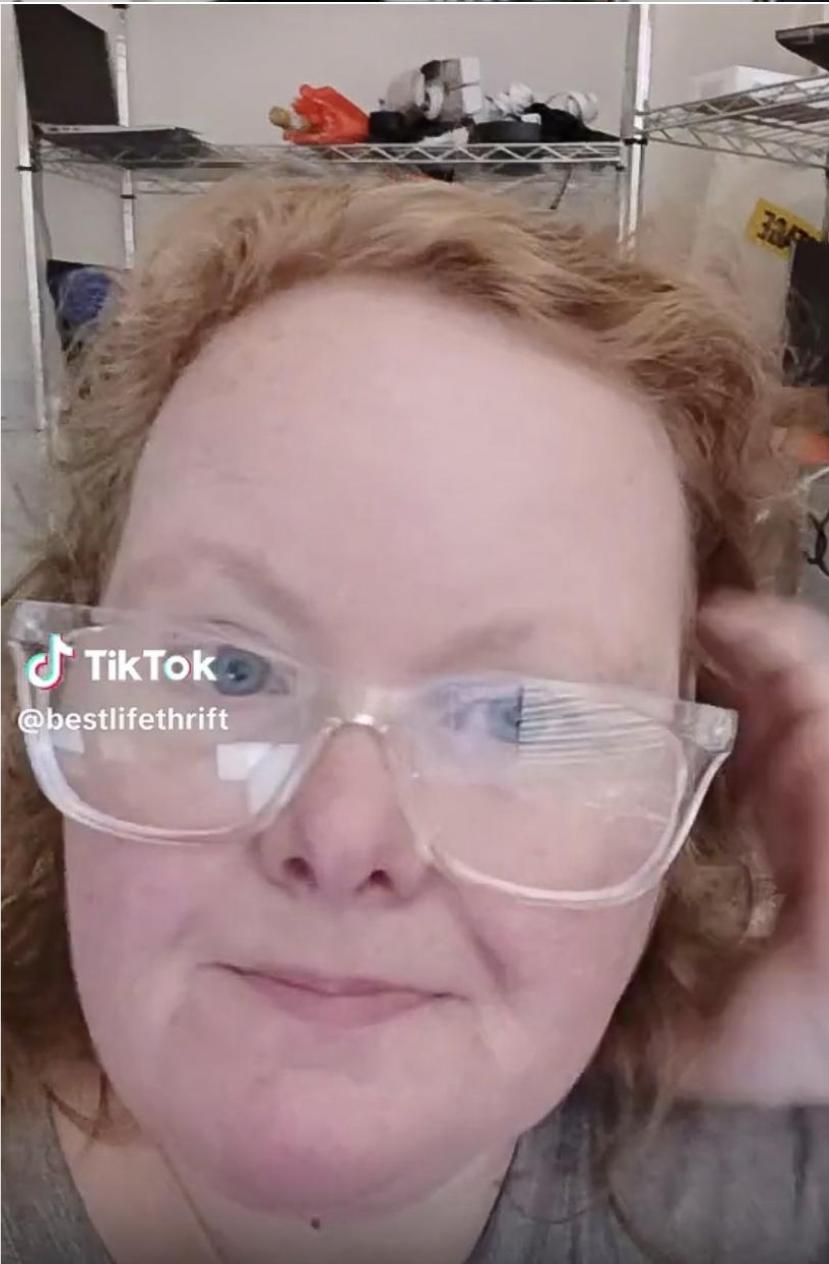


Synthetic Identities



Real time Video Deep fake Face Swap – And It Continues...





Screenshots from Chapman's June 2023 Tiktok video with laptops in the background.

What Does These Organizations Have in Common?

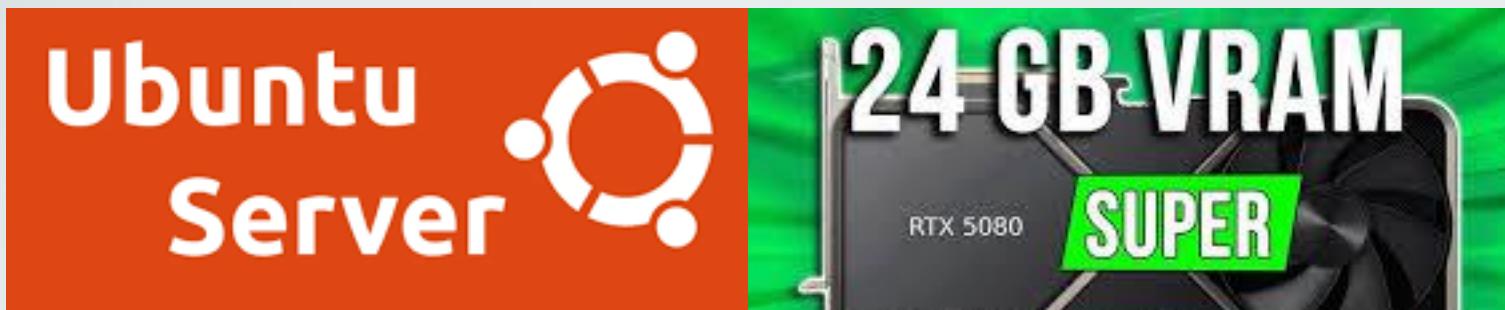
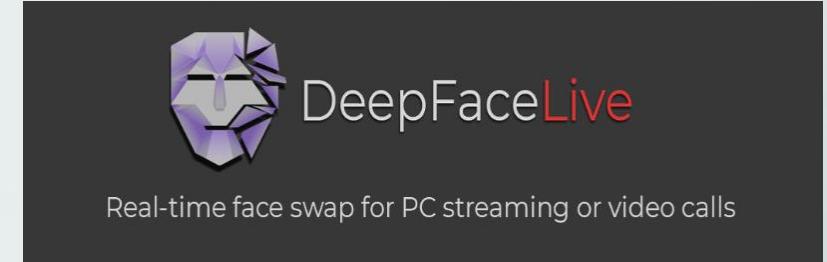


KS &
CER



Create Deepfakes

DEEPFAKE OS



Source: <https://deepfakedashboard.com/deepfakeos>

© 2025 KnowBe4, Inc. All rights reserved.

Deepfake Creation Process

Determine the mark

Identify the target for the deepfake



Source Videos, Audio

Collect necessary media materials



OSINT investigation

Gather information about the target

Upload Audio to Voice Generator

Transfer audio to voice synthesis tool



Download material

Acquire the media files

Upload to Video Generator

Transfer audio to video synthesis tool



Create audio file from text

Generate audio from text input

Present

Showcase the final deepfake



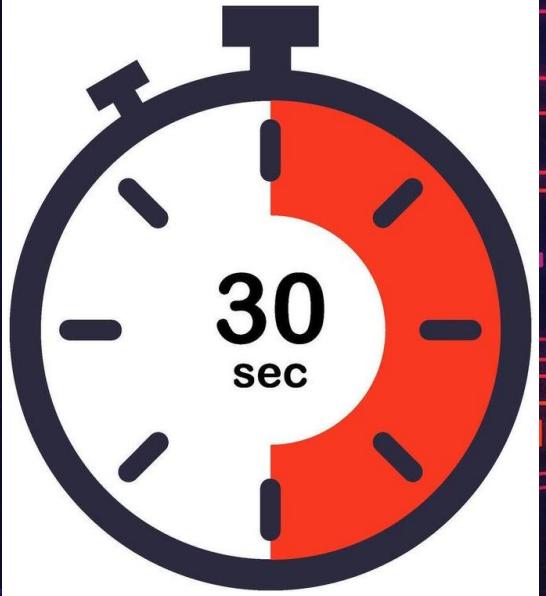
Generate Video

Create the deepfake video

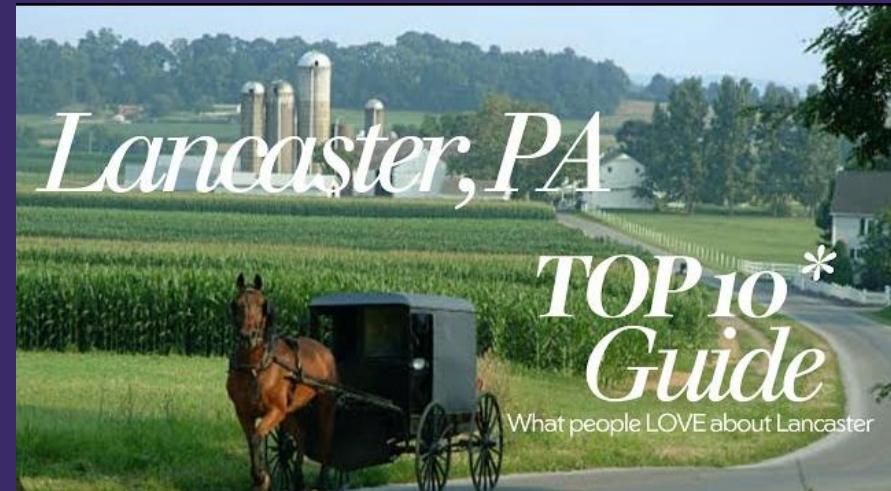


Audio Cloning / Deepfakes

- **Cloning Tools**
 - ElevenLabs, PlayHT, Resemble
- **LLM & Interactive Tools**
 - Dialpad, VAPI
- **Realtime Audio Cloning**
 - Voice.ai, RTVC, Altered.ai
- **Audio Downloaders**
 - YouTube Downloader, Airy, ytmp3.cc
- **Audio Cleaning**
 - Audacity / Audition / Descript



Audio Tips





Video Cloning / Deepfakes

- **Image to Video**
 - Hedra, LemonSlice, Synthesia.io, HeyGen, Deepfake Offensive Toolkit (DOT)
- **LipSyncing (Advanced)**
 - KlingAi, Pixverse, Synclabs, Akool
- **Realtime FaceSwap - Advanced**
 - DeepFaceLab, Deep-Live-Cam, SwapFace, MagicCam
- **Video Generation**
 - VEO3, KlingAI, Pixverse



Video Tips

SOURCE MATERIAL

X Low-Res, Poor Quality

✓ HIGH-RES, WELL-LIT

PODCAST VIDEOS ARE GREAT!

TECH & TRAINING

NVRAM
It's all about the VRAM!

Long Training Hours

DEMONSTRATION

LIVE CISO SACRIFICE OR
COMMODORE 64

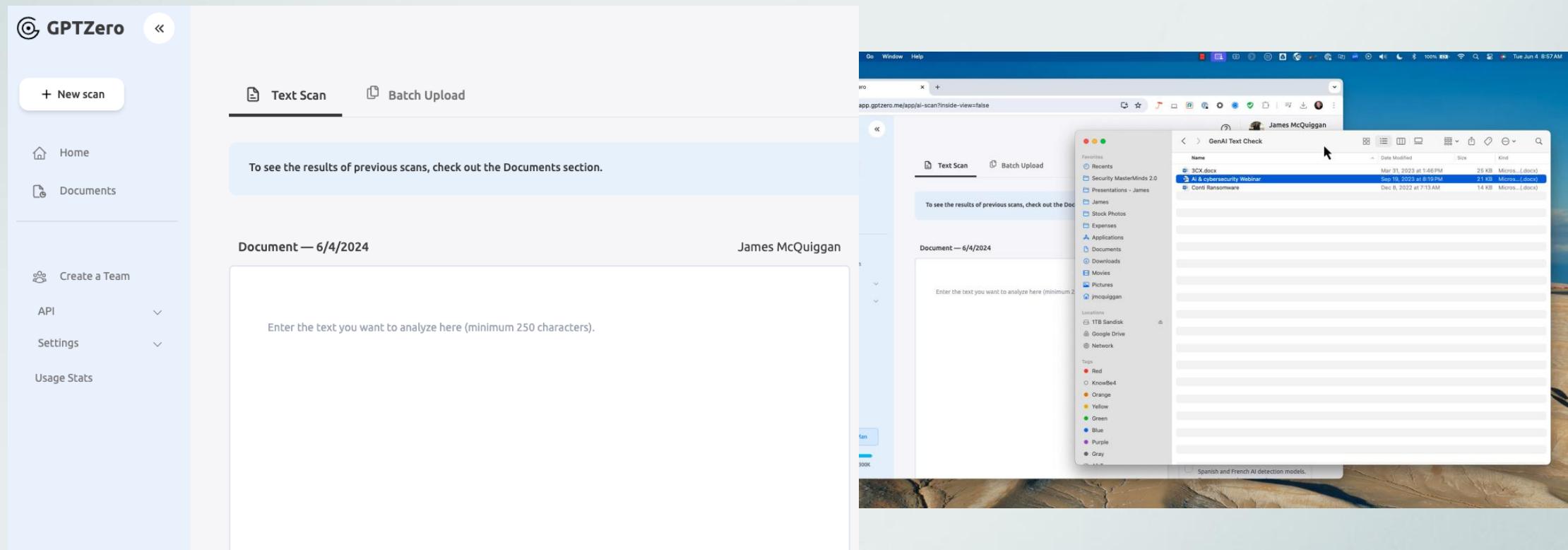
DEMO



Detect Deepfakes

GPTZero – GenAI Synthetic Text Detection

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



Audio File

 detect.resemble.ai/results/0b7e6bac1708987c39e00b3d2805fd0c

Resemble Detect

Detect deepfake audio from any source with our powerful AI Model. [Try it out yourself →](#)





Result: Fake

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

Deepfake Dashboard

The Deepfake Dashboard is a comprehensive platform for threat analysis and reporting. It features a navigation bar with links to Services, Training, Legal, and Account, along with a Psyber Labs logo. The main content area includes a "Showcased Reports" section for the "Mr. Beast iPhone Scam" with a "Threat Level: Moderate" rating and a 3D radar chart. Other sections include a brain visualization, a bar chart for "Top CCVEs", and a scatter plot for "Sophistication" vs "Maliciousness". A footer banner at the bottom reads "Detect | Dissect | Defend".

DEEFAKE DASHBOARD

Services Training Legal Account Psyber Labs

Showcased Reports

Mr. Beast iPhone Scam

you're one of the 10.000 lucky people who'll get an iPhone 15 Pro for just \$2.

Click to View Full Report

Threat Level: Moderate

Top CCVEs

Bias Type	Percentage
Confirmation Bias	40%
Authority	35%
Emotional Load	30%

Distribution Credibility Interactivity Familiarity Evocation

Sophistication Maliciousness

Detect | Dissect | Defend

10 Best AI DeepFake Detector Tools



Synthetic Video Detection Challenges

Lack of Standardization

No uniform method exists for detecting deepfakes effectively.

High False Positives

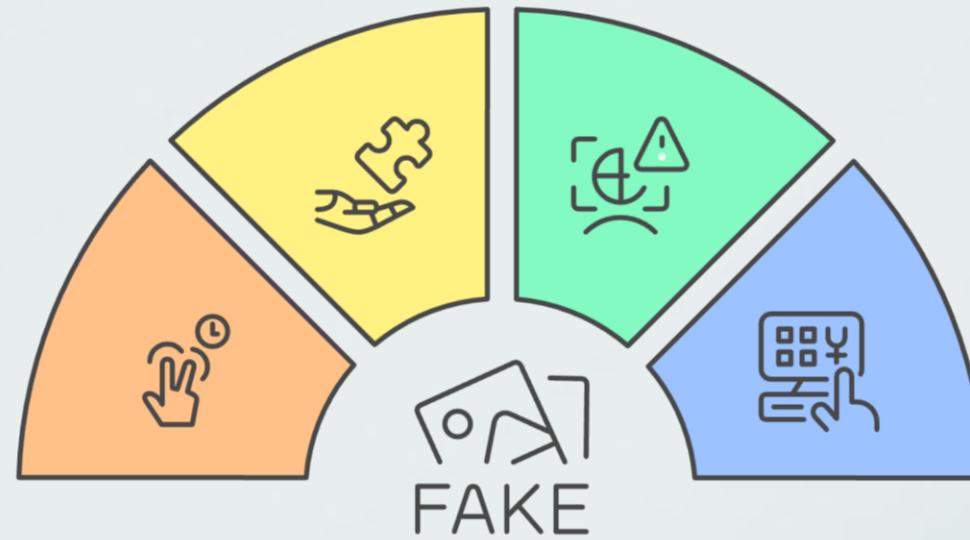
Many legitimate images are incorrectly flagged as deepfakes.

Non-Real-Time Detection

Detection processes do not operate in real-time, delaying responses.

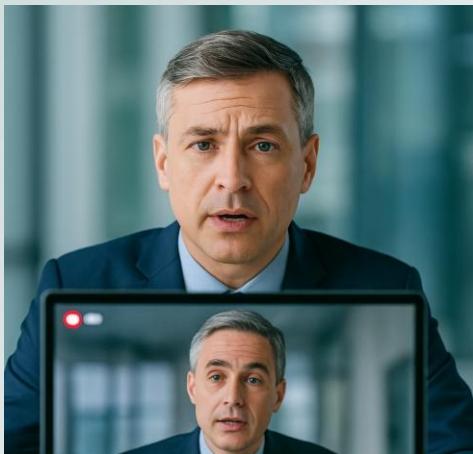
Manual Effort Required

Detection still relies heavily on human intervention and analysis.



Generation technology is outpacing detection technology

Processes / People - Tabletop Exercises



DEEFAKE CEO WIRE TRANSFER SCAM

A video call appears to come from your CEO while traveling abroad. They urgently ask the CFO to authorize a wire transfer for a confidential acquisition.



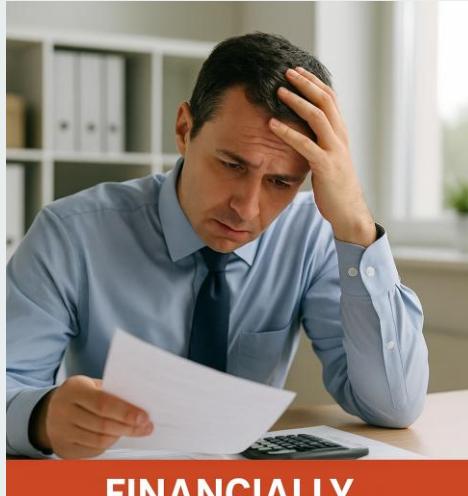
EMPLOYEE FALLS FOR DEEFAKE CEO WIRE TRANSFER SCAM

An employee receives a video call from the supposed CEO, mimicking their appearance and voice. Tricked by the realistic deepfake, they initiate a fraudulent



DEEFAKE PHONE CALL SCAM

You receive a call that appears to be from your manager. Their voice urgently asks for sensitive information to resolve a payroll issue.



FINANCIALLY DISTRESSED BUSINESS SCAM

A struggling small business is pressured to make purchases or pay invoices, falsely hoping that it will improve their finances.

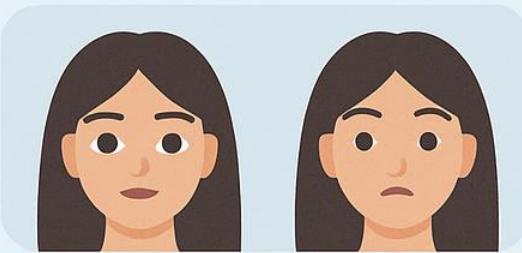


PHISHING SCAM

You receive an email claiming to be from your bank. To prevent your account from being locked out, it urgently tells you to click a link and take action.

People - Ways to Detect Deepfakes

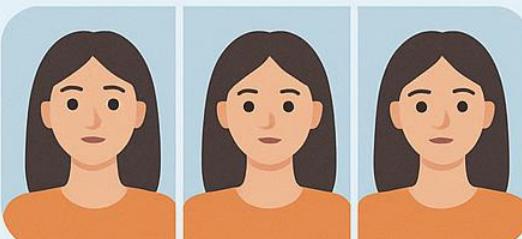
Look for unnatural blinking



Check lighting consistency



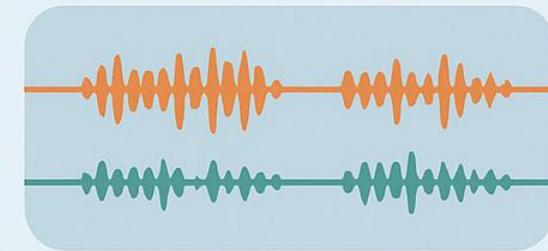
Watch for temporal glitches



Examine edge artifacts



Listen for audio sync issues



Verify through reverse image search



People - What Should We Be Asking?



✗ Is this a deepfake?

✓ Consider these questions...

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

Using Questions / Be Curious

“What is the book
that you
recommended to
me?”

‘I Need to Identify You’: How One Question Saved Ferrari From a Deepfake Scam

- Benedetto Vigna was impersonated on a call using AI software
- Large companies are being increasingly targeted with deepfakes



Defense in Depth for AI & Social Engineering

MFA

- Use MFA Wherever Possible - Non-phishable MFA too
- Avoid SMS and verify all requests that you didn't initiate

Email Auth

- DMARC / DKIM / SPF to prevent email spoofing!
- Only 15% of orgs use this.

AI vs AI

- Use AI with your cybersecurity gateways & products
- Behavioral Analysis in email and users

Zero Trust

- Be mindful of your inbox. Be skeptical, if you're not expecting it and it's a strange or unusual request

Training

- Frequently educate and assess your users
- Leverage threat intelligence for your threat landscape

**People are a
critical layer within
security programs**





TRUST
& VERIFY

BE
SKEPTICAL



POLITELY
PARANOID



Questions



Final Words

- Wrap-up



ACTION PLAN



People (Security Awareness & Training)

- Human Risk Management Program
- Conduct deepfake awareness training for all employees
- Implement verification protocols for executive or financial requests
- Use security questions in high-risk communications



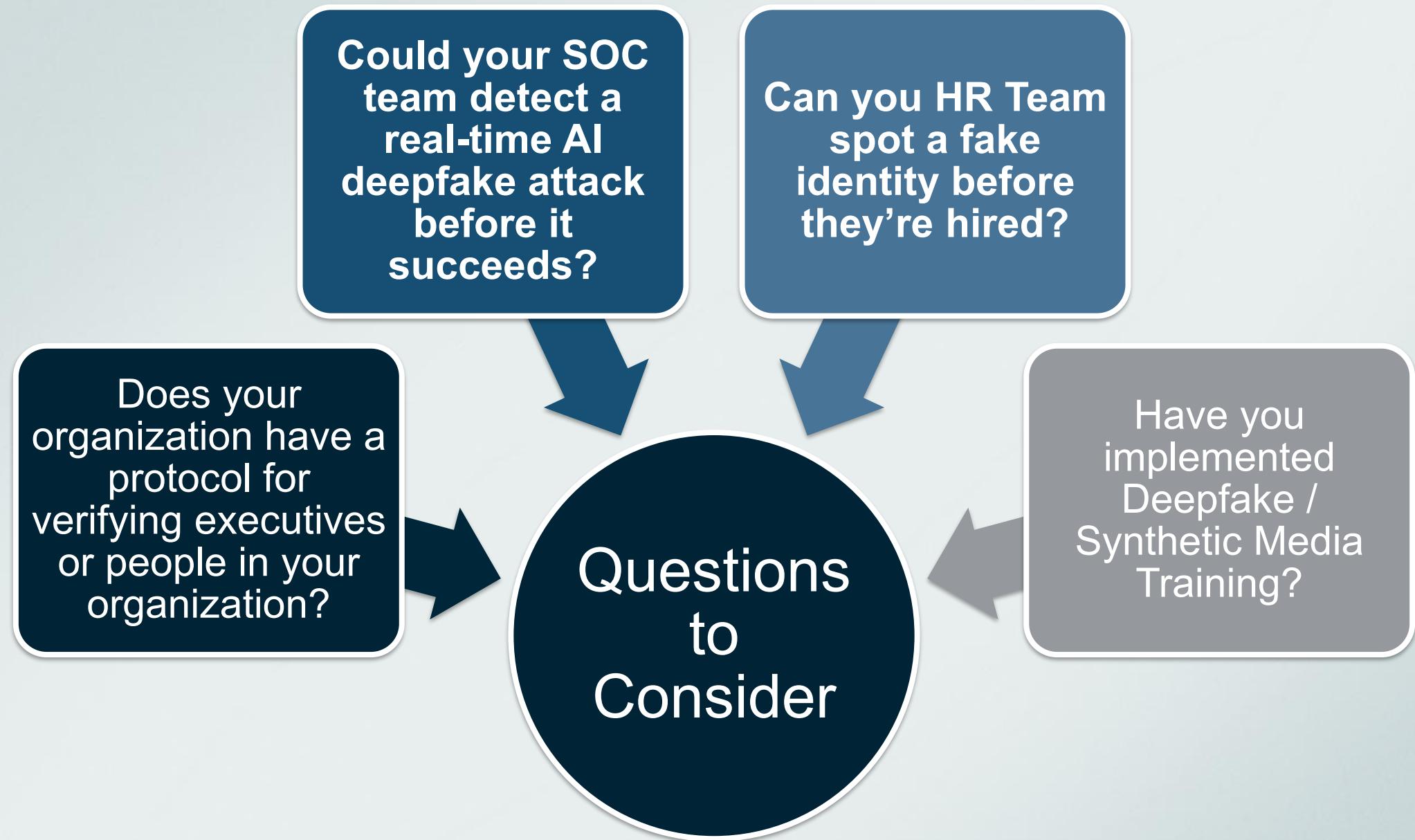
Processes

- Ask location-based verification questions
- Require cameras on for remote interviews with assessments
- Establish a SOC protocol for real-time AI threat detection
- Table Top Exercises Assessing Deepfake Scenarios



Technology

- Deploy AI-driven deepfake detection software
- Implement audio and video authentication measures
- Leverage POCs for deepfake detection services / software



Interested in Learning More?

The screenshot shows the RESEMBLE.AI Deepfake Incident Database. At the top, there's a navigation bar with links for AI Voice Generator, Detect, Resources, Government, Pricing, and Sign In. Below the navigation is a section titled "AI Safety" with a green background. A large title "Deepfake Incident Database" is centered. To its right is a detailed description of the database: "A curated database tracking verified incidents where deepfake technology has been used to target specific individuals or organizations. This collection documents cases of AI-generated synthetic media being weaponized for impersonation, manipulation, or deception, providing insights into the real-world impact and evolution of deepfake threats." Below this text is a button labeled "Learn More About Detection".

<https://www.resemble.ai/deepfake-database/>

The screenshot shows the AIAAIC Repository. At the top, there's a header with the text "The independent, open, public interest resource detailing incidents and controversies driven by and relating to AI, algorithms and automation. [More](#)". Below the header is a circular image of a complex mechanical or electronic device. To the right of the image is a section titled "Latest entries" which lists several news items. Further down is a section titled "Recent updates" which also lists news items. The overall design is clean and modern.

<https://www.aiaaic.org/aiaaic-repository>

The screenshot shows the AI Incident Database. The left sidebar contains a navigation menu with options like Discover, Submit, Welcome to the AID, Discover Incidents, Spatial View, Table View, List View, Entities, Taxonomies, Submit Incident Reports, Submission Leaderboard, Blog, AI News Digest, Risk Checklists, and Random Incident. The main content area features a search bar with the placeholder "Search over 3000 reports of AI harms" and a "Search" button. Below the search bar is a "Discover" button. A specific incident report is highlighted: "Incident 950: NullBulge's AI-Powered Malware Allegedly Compromises Disney Employee and Internal Data". The report summary includes a link to "A Disney Worker Downloaded an AI Tool. It Led to a Hack That Ruined His Life." and a timestamp "waj.com 2025-02-26". There are also "Read More" buttons and navigation arrows.

<https://incidentdatabase.ai>



**Education,
Preparation,
and healthy
security habits
are the only defense**

Most Secure Woman?



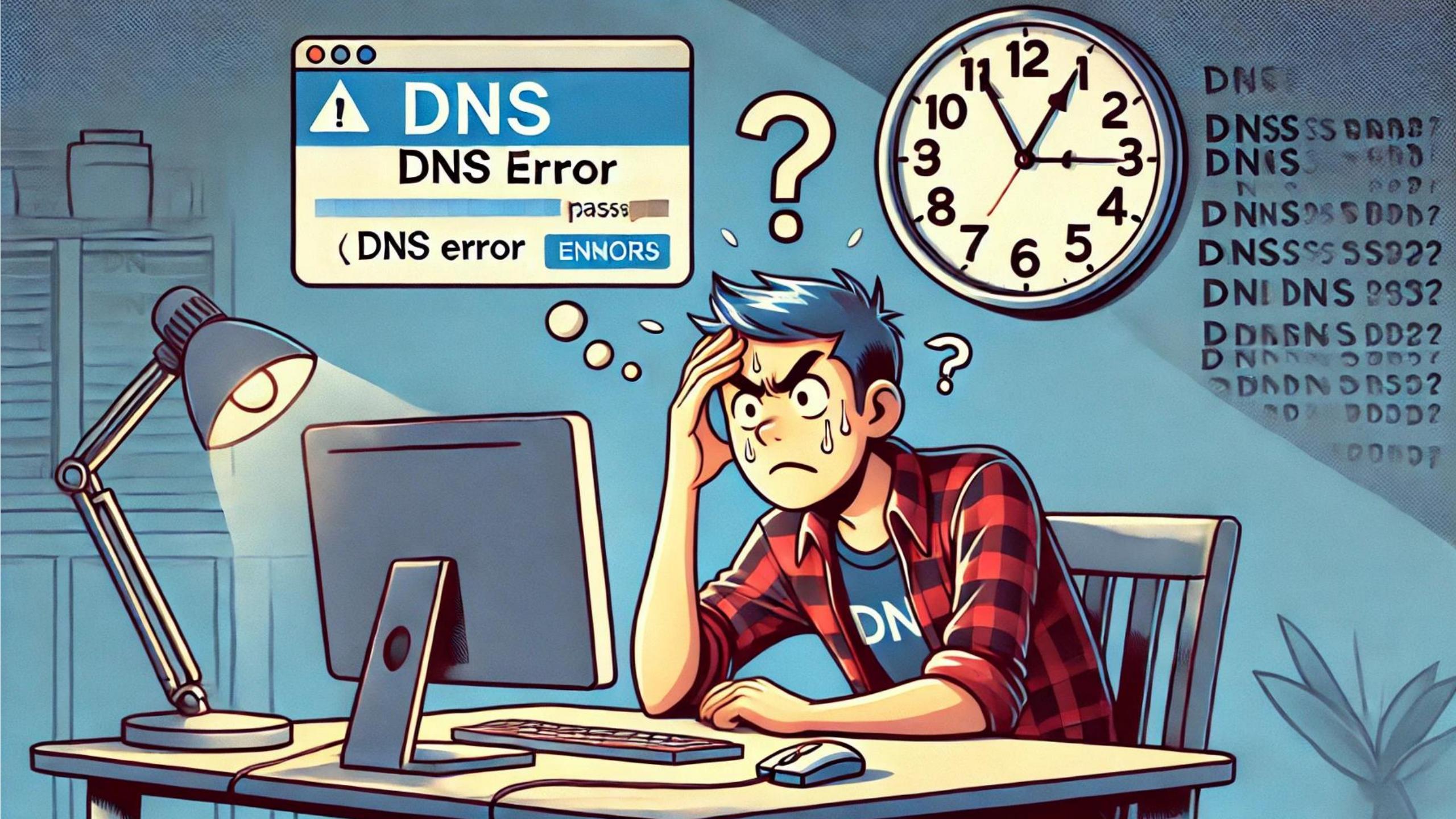
MFA



Emma Faye



Multifactor
Authentication



Yes... I have a
way of keeping
track of my
Dad Jokes



I Have a Favor to Ask You

jmcquiggan@knowbe4.com



<https://talk.ac/jamesmcquiggan?code=TALK>

