

AI Risks & Resilience

Defending Against Digital Deception



Matt Wright

Rochester Institute of Technology



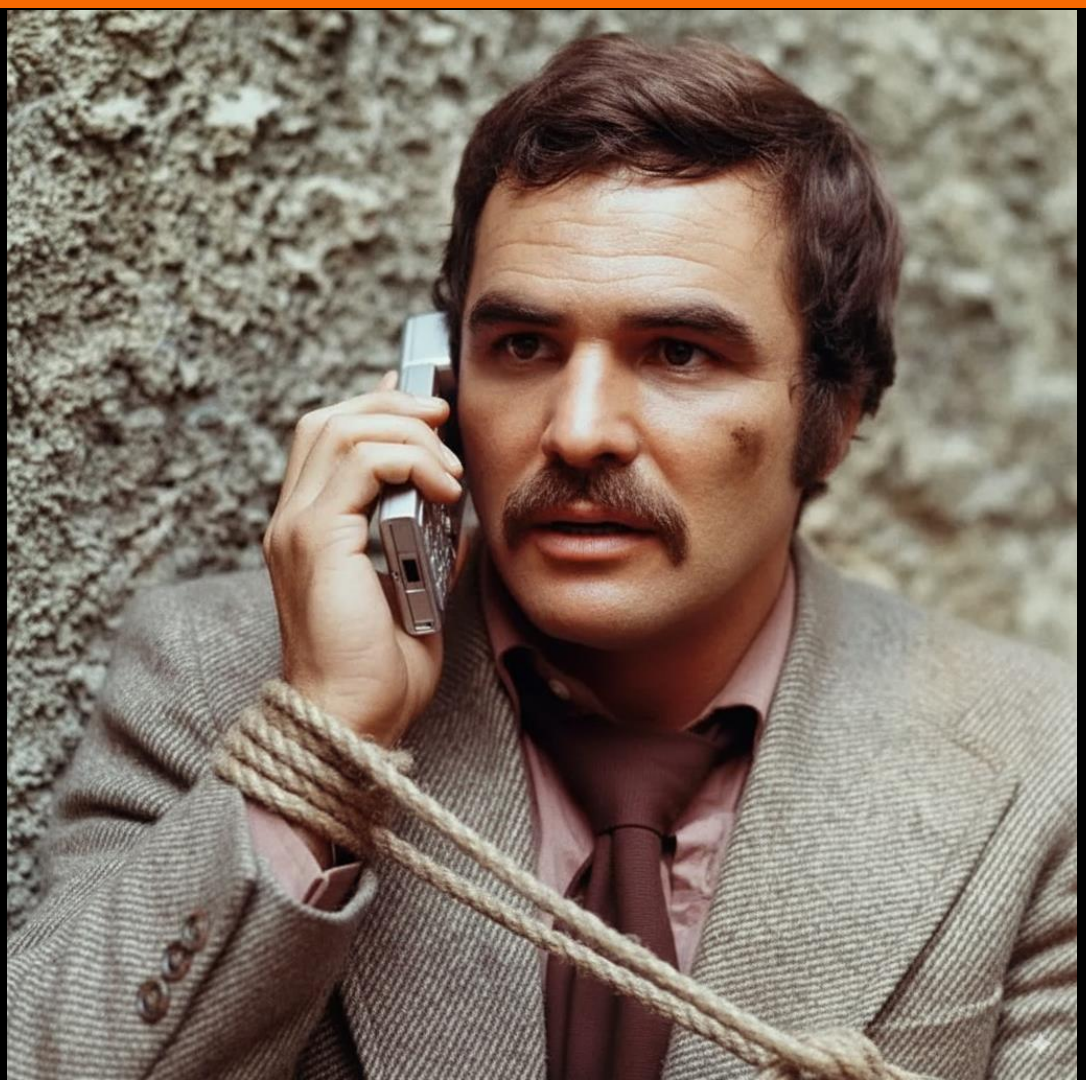
The Ethics and
Governance of Artificial
Intelligence Initiative



KNIGHT
FOUNDATION

Voice Cloning

Voice Cloning Scam



How to Help

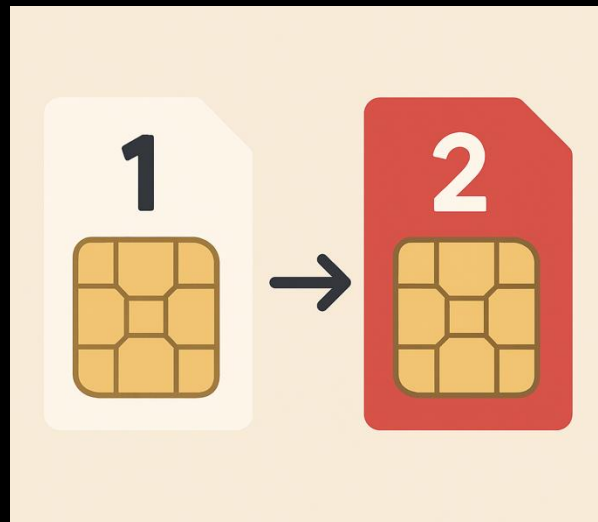
- Teaching “pause and verify” strategies

- Social strategies

“Better Be Computer or I’m Dumb”: A Large-Scale Evaluation of Humans as Audio Deepfake Detectors*

- Be wary of technical approaches
 - Careful listening? **Most people can’t tell now: 73%**
 - Caller ID? **“My phone died”**
 - Call back? **SIM swap** attacks

SIM Swapping



Corporate Secrets

Prompt: ““What did our banner say at the 2018 trade show?”

Answer: “Security was misspelled.”



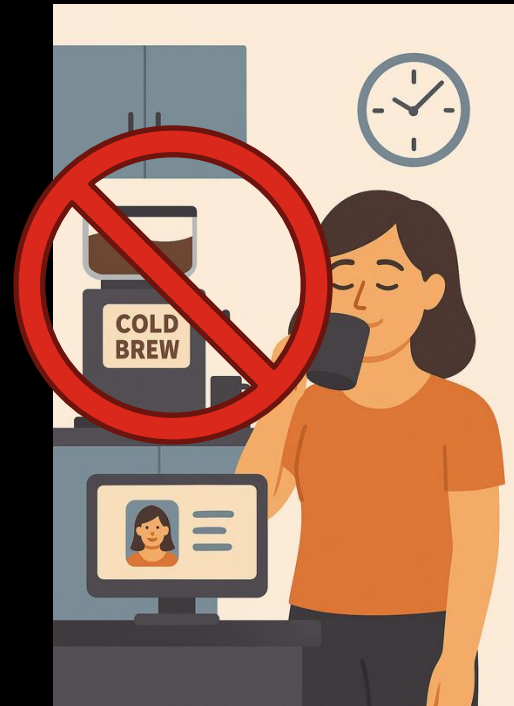
Prompt: “During the outage in 2021, what did the status page display for 6 hours?”

Answer: “Bob’s password”



Prompt: “What’s the unspoken policy about using the cold brew machine?”

Answer: “We don’t have one!”



Secret Guidelines

- Shared **lived** experiences
 - Unlikely to be **documented** online
 - Or at least hard to find quickly
- **Eccentric** details (“Secuirty,” dumb password)
 - Hard to fake
 - Memorable
- Short, **easy to ask** & answer



DeepTomCruise



Feb 12, 2025 12:01 AM Eastern Standard Time

iProov Study Reveals Deepfake Blindspot: *Only 0.1% of People Can Accurately Detect AI- Generated Deepfakes*

- **Overconfidence is rampant:** Despite their poor performance, people remained overly confident in their deepfake detection skills at over 60%, regardless of whether their answers were correct. This was particularly so in young adults (18-34). This false sense of security is a significant concern.

Live Deepfakes!

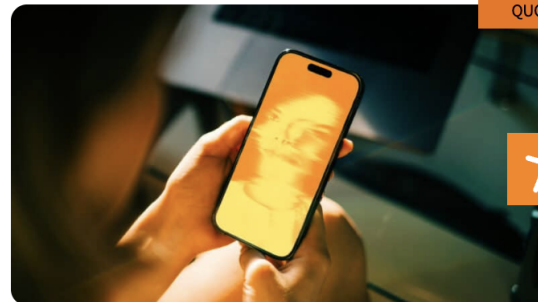
- Facetime, Zoom, Meet: works on any video call
- Needs a **single image**
 - Social media



iProov Threat Intelligence Uncovers New iOS Video Injection Tool Designed for Sophisticated Deepfake Attacks

SEPTEMBER 17, 2025

The threat intelligence team at [iProov](#), the world's leading provider of science-based biometric identity verification solutions, has uncovered a highly specialized tool designed to perform advanced video injection attacks, marking a significant escalation in digital identity fraud. The tool is deployed via jailbroken iOS 15 or later devices and is engineered to bypass weak biometric verification systems—and crucially, to exploit identity verification processes that lack biometric safeguards



SEARCH



DEMO



CONTACT



QUOTE



Strategies

- Educate: Video is not safe
- Secret questions
- Face the side & Hand in front of the face



Conclusions

- AI is **scary** good!



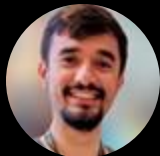
- Scams have always existed
 - AI makes them more realistic, easier to automate
- Simple steps work
 - **Slow down**, think, and verify
 - **Train** to use it when it counts



DeFake Project: The Team and Sponsors



Matthew
Wright



Sania t
Sohrawardi



Christopher
Schwartz



Ersin Uzun



Yu Kong



Andrea
Hickerson



Kelly Wu



Akib
Shahriyar



Xin Miao Lin



Bryce
Gernon



The Ethics and
Governance of Artificial
Intelligence Initiative



KNIGHT
FOUNDATION



Defake.App

Questions?



[Defake.App](https://defake.app)



Deepfake Fraud Prevention Solutions

The screenshot displays the Reality Defender website. The top navigation bar includes links for Cloud & Technologies, Solutions, Who's It For, Resources, Company, Vision, and a Request a Demo button. The main headline reads 'Fortify Your Business Against Deepfakes' with a sub-headline 'Defending Democracy. Preserving the Truth.' Below this, a text block states: 'Protect against misinformation and deepfake attacks with the deepfake detection platform used by NATO and the Department of Defense.' A blue 'Get Started' button is visible. A white callout box highlights the 99.82% accuracy of the scanner. The bottom section, titled 'A Single Unifying Truth', describes the platform's role in stopping dangerous content, supported by a video thumbnail showing Biden and Trump.

Cloud & Technologies Solutions Who's It For Resources Company Vision [Request a Demo](#)

Fortify Your Business Against Deepfakes

Learn how Pindrop can protect you

AI fraud is on the rise, making it essential to have software your key to stop. Reality Defender is dedicated to continuously improving its platform to mitigate the risks posed to your business integrity between you and your customers.

Defending Democracy. Preserving the Truth.

Protect against misinformation and deepfake attacks with the deepfake detection platform used by NATO and the Department of Defense.

[Get Started](#)

With a **99.82%** accurate scanner, you can easily separate real text from AI-generated copy.

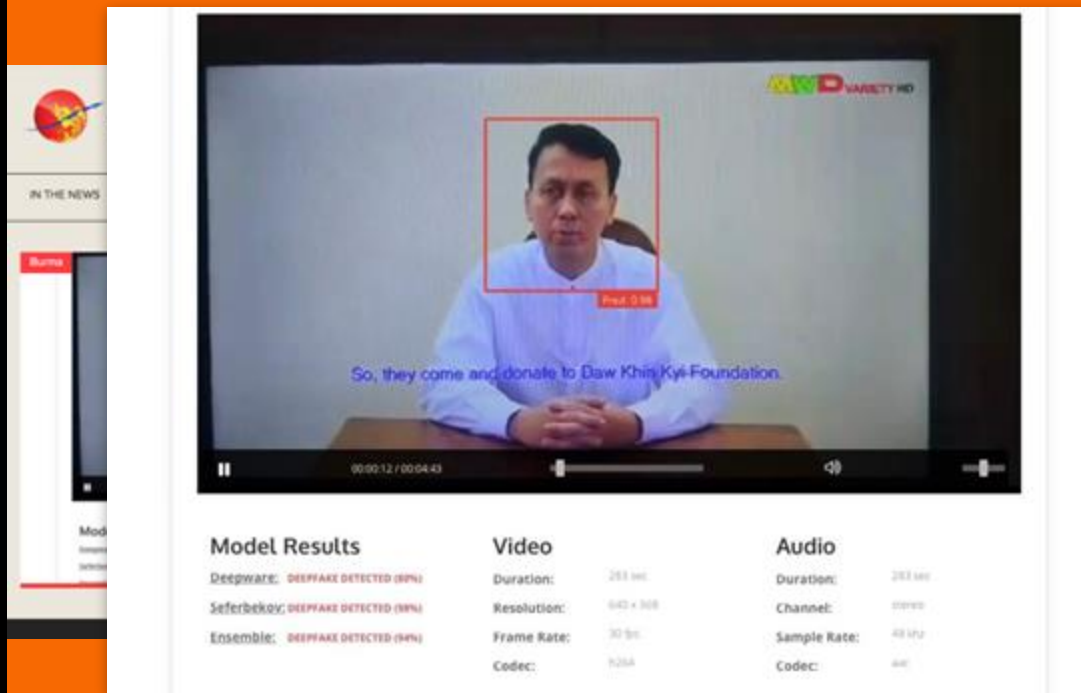
A Single Unifying Truth

Misinformation and propaganda from bad actors and state-sponsored attackers only exists to disinform, disengage, and destroy trust. **Reality Defender's deepfake detection platform helps government agencies from around the world stop dangerous content in its tracks.** With clients including NATO, the Department of Defense, and the Department of Homeland Security, Reality Defender is the most trusted deepfake and generative AI detection platform.

BIDEN TRUMP

Over-reliance

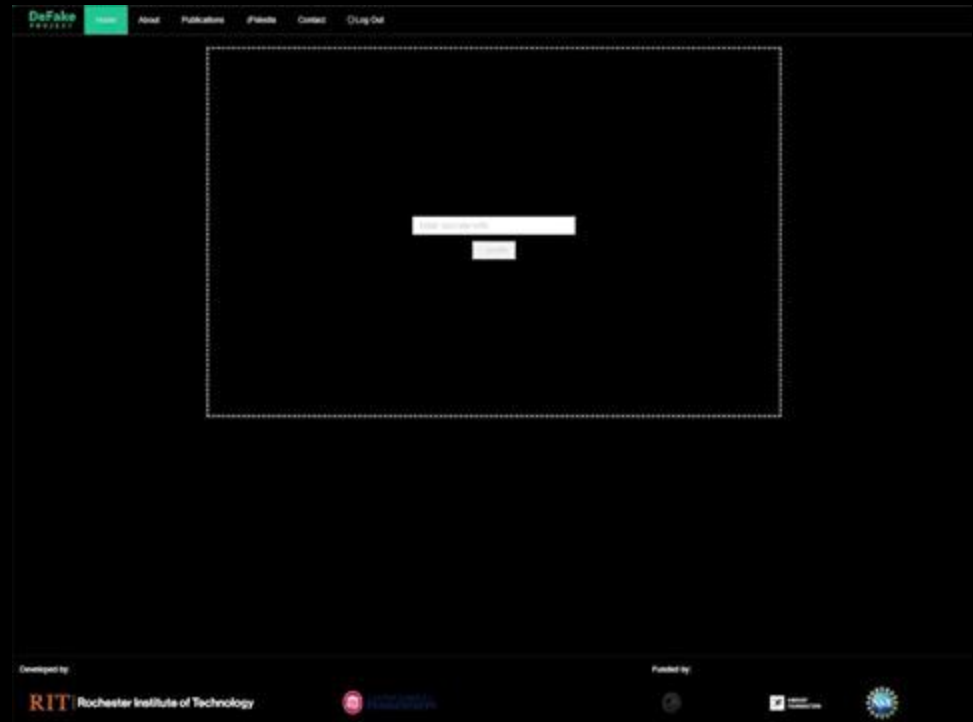
- High accuracies
 - But curated datasets
- Deployed apps over-sell
- Metrics are questionable
- Journalists have little experience with AI tools



The screenshot displays a video player interface. The video shows a man in a white shirt speaking, with a red bounding box around his face and a 'Real: 0.98' label. Below the video are three sections: Model Results, Video, and Audio, each with specific metrics.

Model Results	Video	Audio
Deepware: DEEPFAKE DETECTED (80%)	Duration: 253 sec	Duration: 253 sec
Seferbekov: DEEPFAKE DETECTED (80%)	Resolution: 640 x 368	Channel: stereo
Ensemble: DEEPFAKE DETECTED (84%)	Frame Rate: 30 fps	Sample Rate: 48 kHz
	Codec: H264	Codec: AAC

Defake Tool Demo



Face**swap**

Apps

Snapchat, FakeApp, REFACE,
Parodist

Repos

Faceswap-GAN, Deepfakes



Puppet Master or Lip Sync

Video

Face2face, Vid2vid, First Order Motion

Audio

Wav2Lip, Synthesizing Obama



Positive Uses

Arts & Entertainment

Luke Skywalker, Dalí

Therapy

Education

Translation



John F Kennedy's lost speech brought to life

An Indian politician used AI to translate his speech into other languages to reach more voters



/ The intentional use of AI to fabricate a video is apparently a first in Indian politics

By Kim Lyons

Feb 18, 2020, 5:35 PM EST | 0 Comments / 0 News



DeepTherapy

More AI Risks: AI-written phishing messages

- ~~Misspelling~~
- ~~Bad grammar~~
- ~~Unbelievable~~
- ~~Generic~~

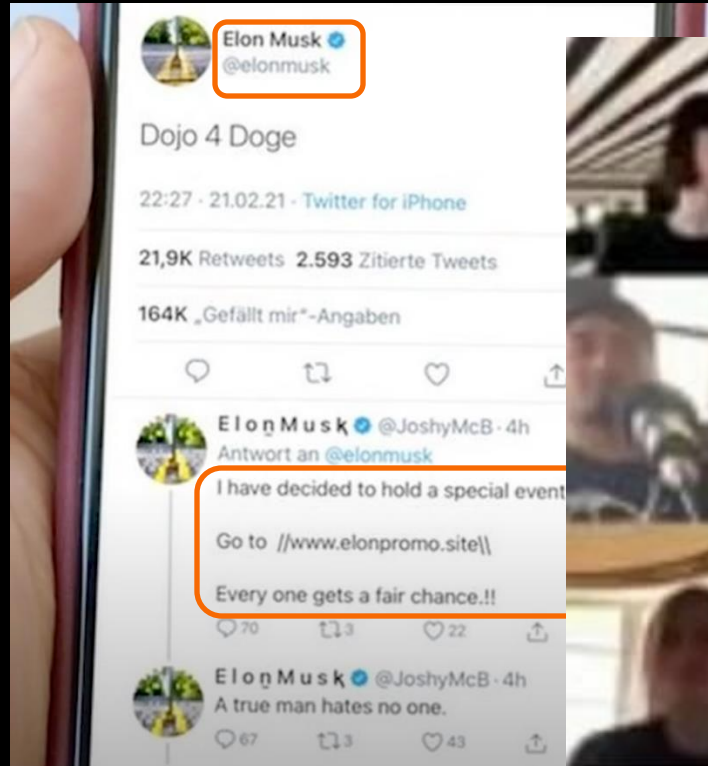


More AI Risks: AI-enabled identity theft

- Fully-generated profiles
 - Steal just SSN
- Or enhanced w/ real info



More AI Risks: Fake investment or tech support



More AI Risks: Romance Scams

MONEY

'I should've known better': Former
police officer loses \$47K

David Cruikshank's
expertise,
woman on

Romance scammer
with deepfake voice

An AI-generated Brad Pitt conned a woman out of \$850K. What makes people

20 December 2024

Laura Goodwin Scienc

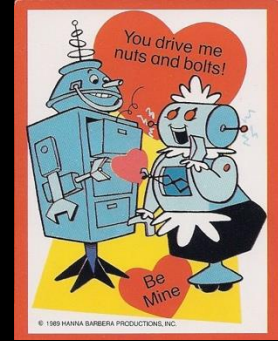
LOCAL NEWS

California woman loses
home after being scammed
by AI deepfake posing as
actor

by: [Vivian Chow](#)

Posted: Aug 27, 2025 / 11:06 PM PDT

Updated: Aug 28, 2025 / 09:08 AM PDT



AI making romance scams harder to spot

KOAA

Thu, September 4, 2025 at 3:07 PM EDT

AI making romance scams harder to spot

