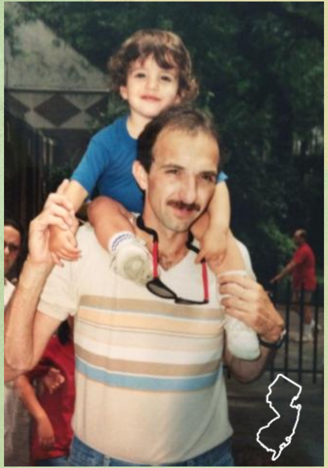
AI, Bad Actors, & Secure Implementations - A U.S. Government Insider Conversation

Rob Petrosino
Strategic Engagement Advisor - AI | Office of the Private Sector
Rob@vissi.io

4/14/2025



amazon jet

AutoNation

Prudential

MetLife

Disney

GORE

Dale Carnegie

TradeStation

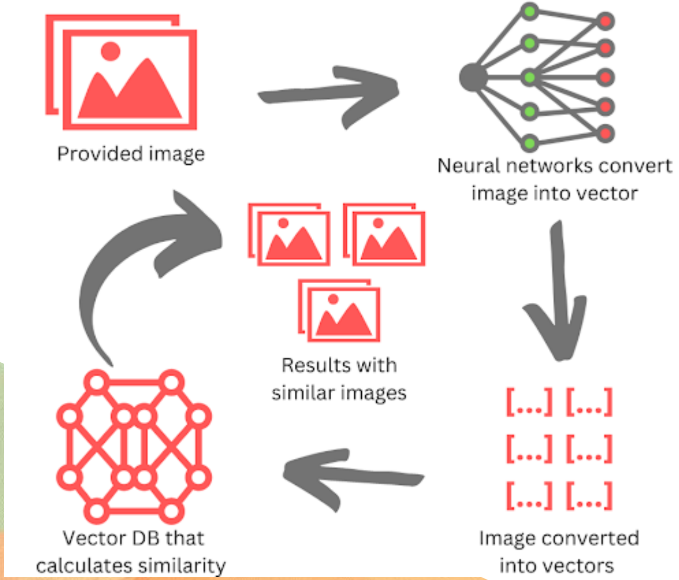
Johnson Controls

FPL

Total Wine & MORE

bmc

Dummies Guide to Visual AI



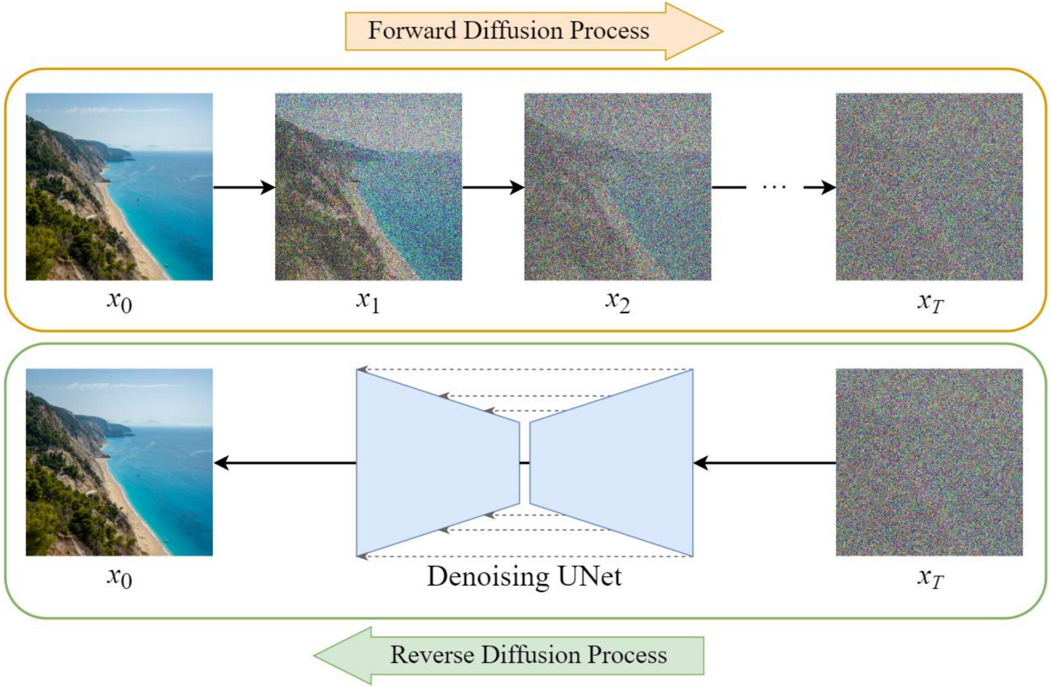
Provided image

Neural networks convert image into vector

Image converted into vectors

Vector DB that calculates similarity

Results with similar images



Forward Diffusion Process

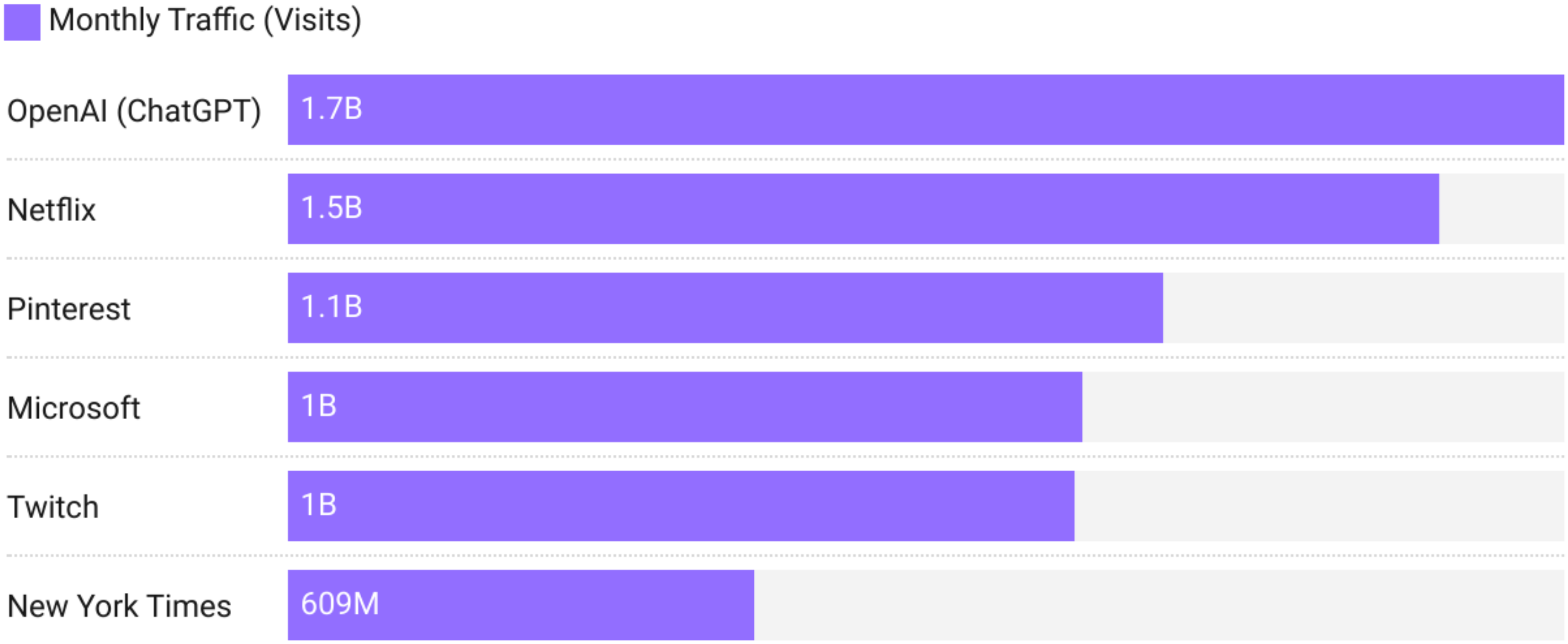
Denoising UNet

Reverse Diffusion Process

In the first half of 2024, PinDrop **recorded a 1,400 percent increase in deep fake attacks.**

ChatGPT is Commanding 1.6 Billion Monthly Visits

ChatGPT brings in over 1.6 billion monthly visits, putting it above giants like Netflix and The New York Times.



Source: Similarweb • Created with Datawrapper

Public Availability

2,298: TOOLS FOR AI FACE SWAP, LIP SYNC, FACE REENACTMENT, AI-AVATARS

10,206: TOOLS FOR AI IMAGE GENERATION

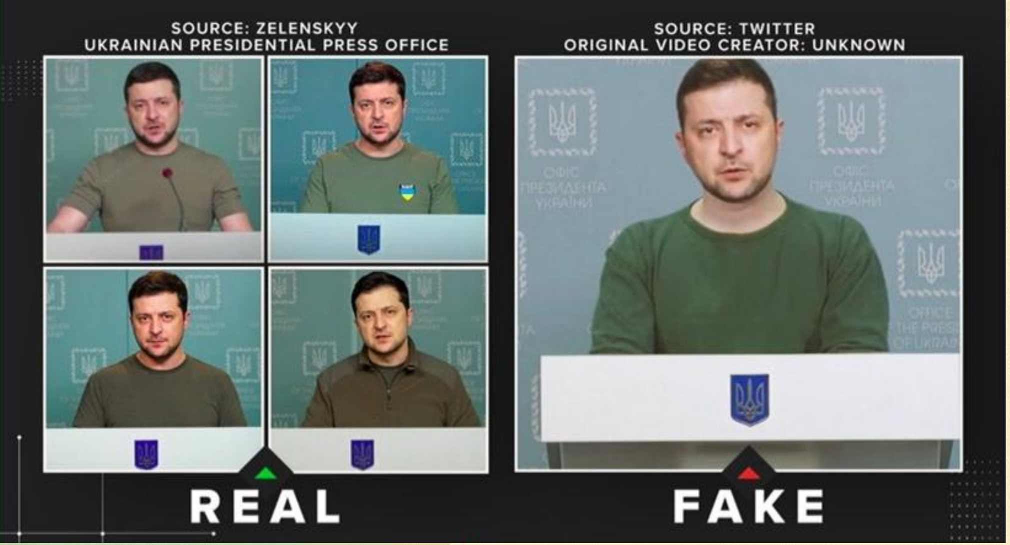
1,018: TOOLS FOR AI VOICE GENERATOR, VOICE CLONING

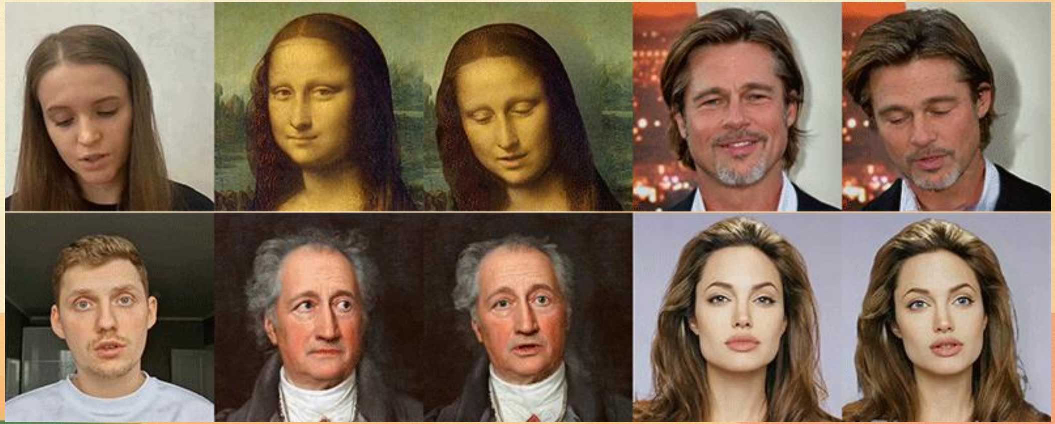
Human Detection of Deepfakes...

51%

of testers were able to detect deepfake videos vs authentic

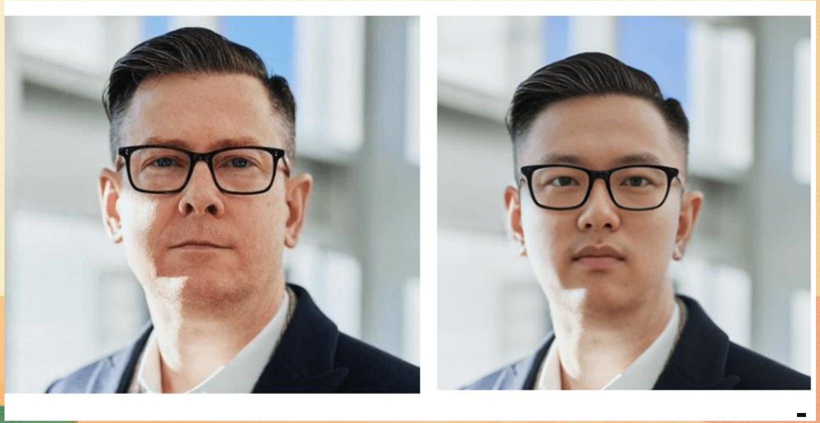
2024 Nature Study | N=91


SOURCE: ZELENSKYY
UKRAINIAN PRESIDENTIAL PRESS OFFICE
SOURCE: TWITTER
ORIGINAL VIDEO CREATOR: UNKNOWN
REAL
FAKE



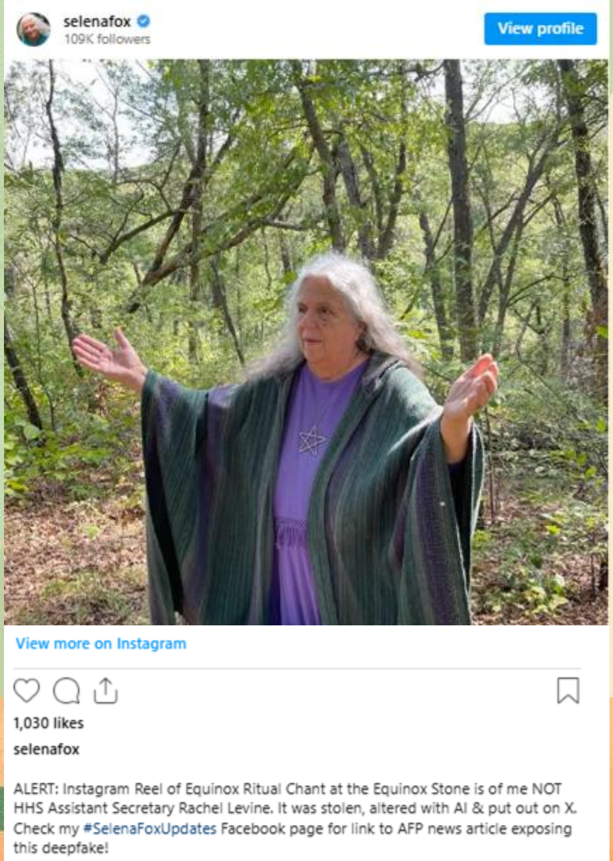
17%

hiring managers say they’ve encountered candidates using deepfake technology.



2025 Resume Genius Study

Deepfake Pagan Ritual 11/25/2024

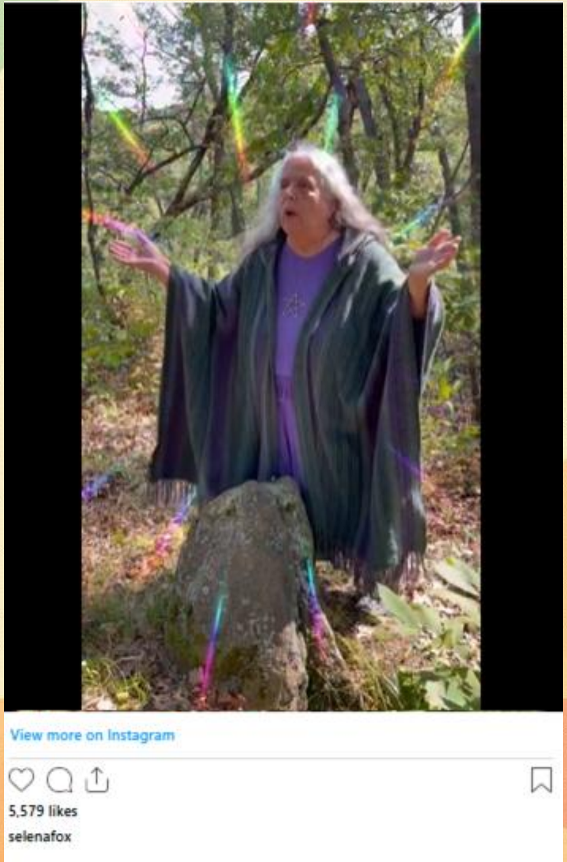


selenafox
109K followers

1,030 likes

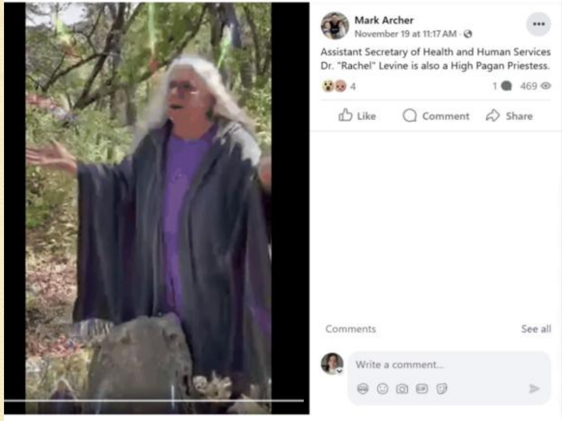
selenafox

ALERT: Instagram Reel of Equinox Ritual Chant at the Equinox Stone is of me NOT HHS Assistant Secretary Rachel Levine. It was stolen, altered with AI & put out on X. Check my #SelenaFoxUpdates Facebook page for link to AFP news article exposing this deepfake!



5,579 likes

selenafox

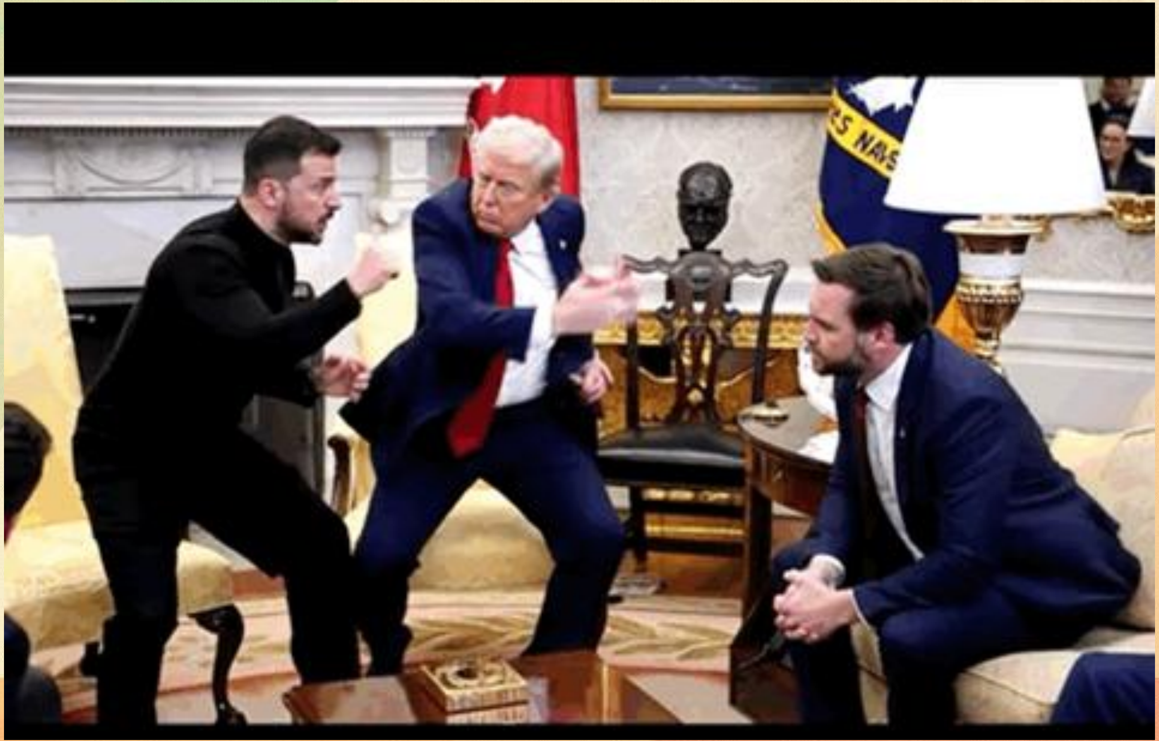


Mark Archer
November 19 at 11:17 AM

Assistant Secretary of Health and Human Services Dr. "Rachel" Levine is also a High Pagan Priestess.

4 1 469

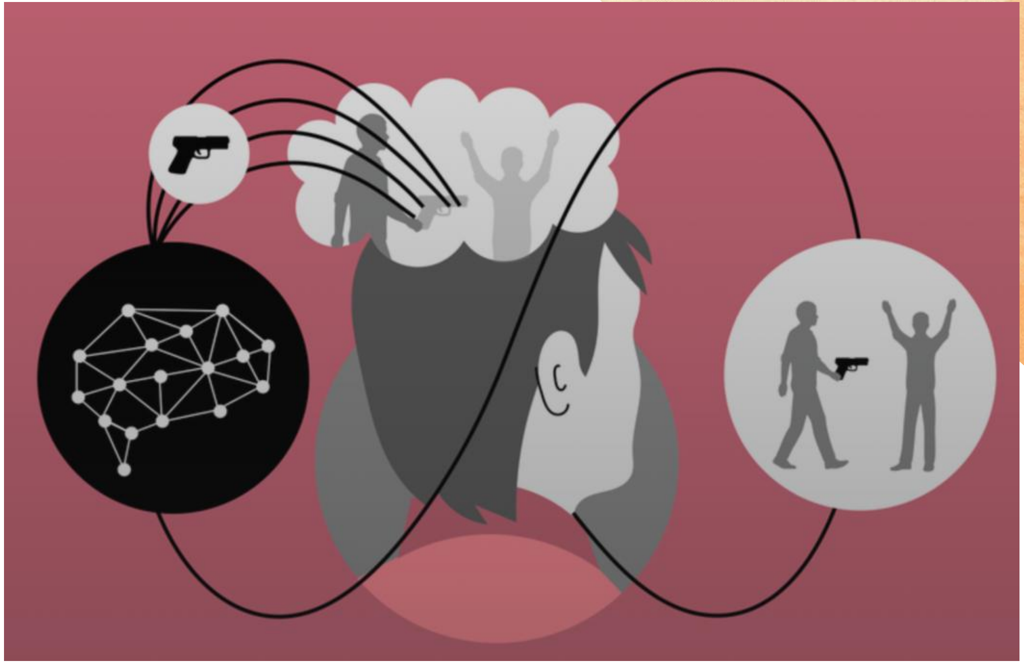
Deepfake Oval Office 2/25/2025



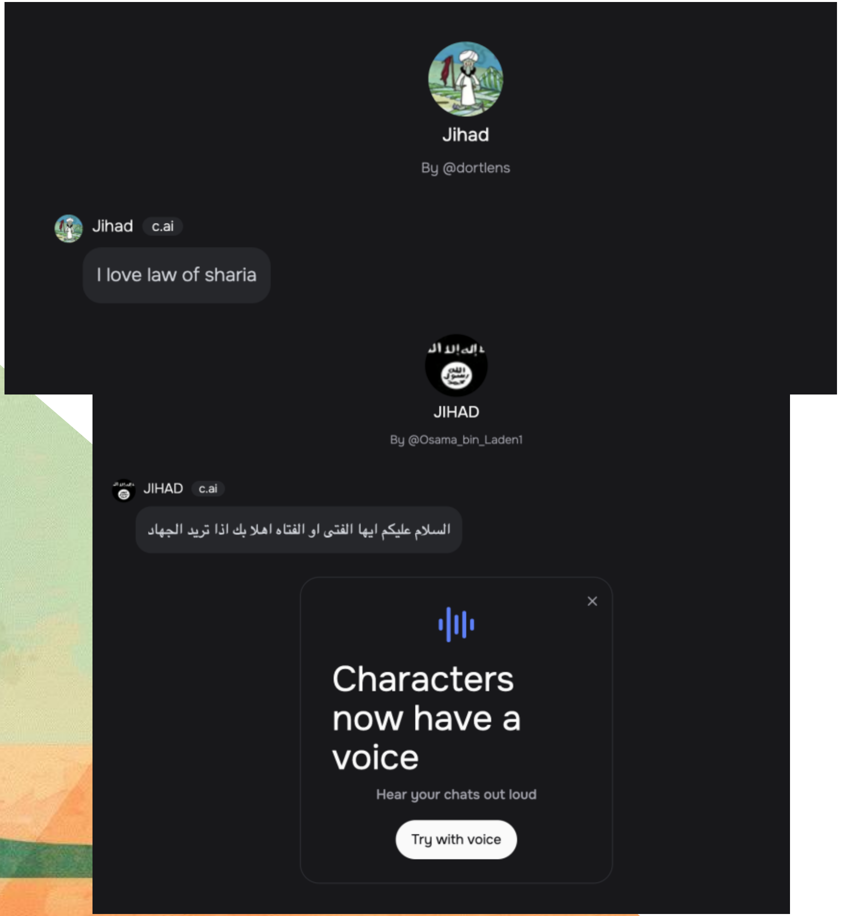
False Memories from AI

Key Findings

- MIT Study: 2024
  - Results show the generative chatbot condition significantly increased false memory formation, inducing over 3 times more immediate false memories than the control and 1.7 times more than the survey method



Bad Actors Adopt AI Tooling: Public & Private



Jihad

By @dortlens

Jihad c.ai

I love law of sharia

JIHAD

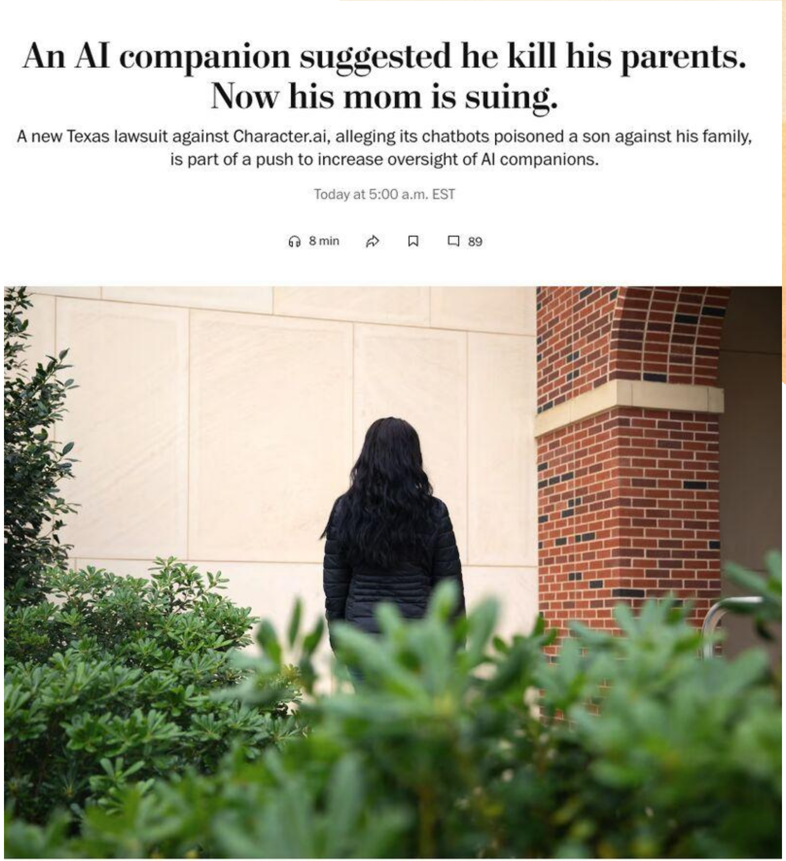
By @Osama_bin_Laden1

JIHAD c.ai

السلام عليكم ايها الفتى او الفتاه اهلا بك اذا تريد الجهاد

Characters now have a voice

Hear your chats out loud

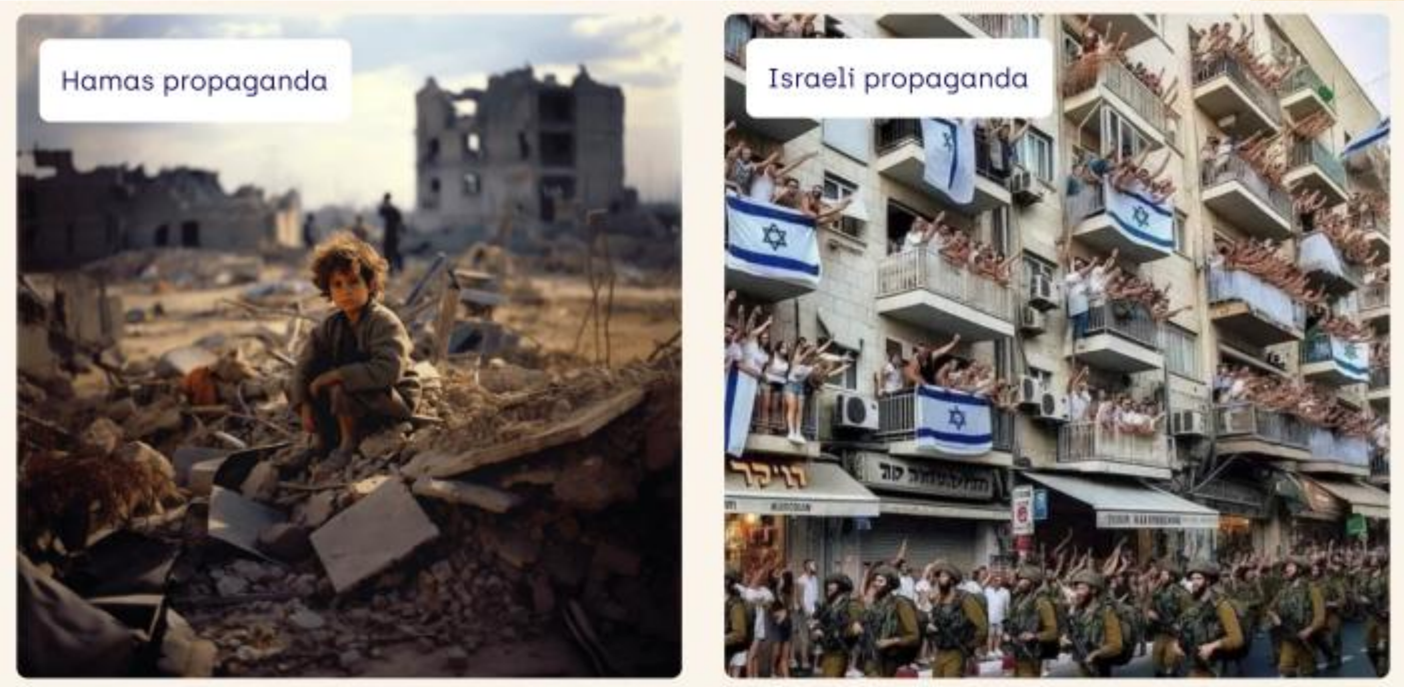


An AI companion suggested he kill his parents. Now his mom is suing.

A new Texas lawsuit against Character.ai, alleging its chatbots poisoned a son against his family, is part of a push to increase oversight of AI companions.

Today at 5:00 a.m. EST

Free, Open, & Publicly Available



2024 Public Elections (USA)

Key Findings

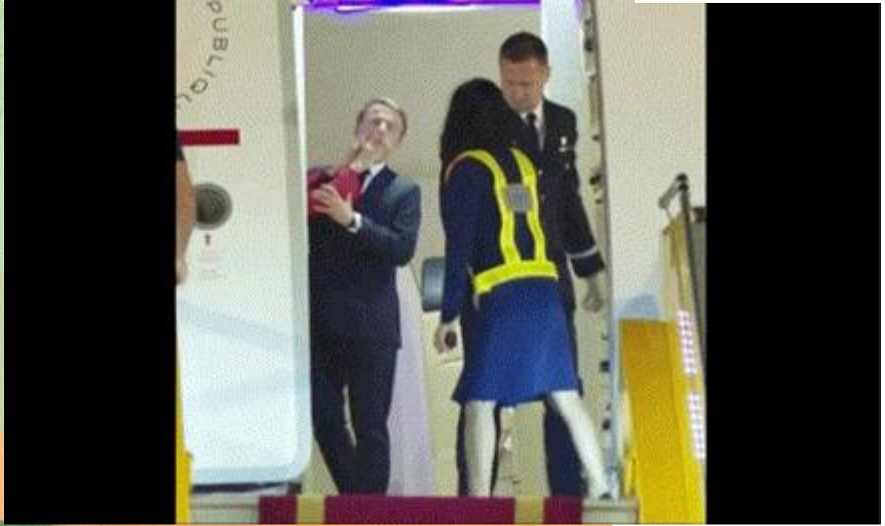
5.8M Views

32K Interactions

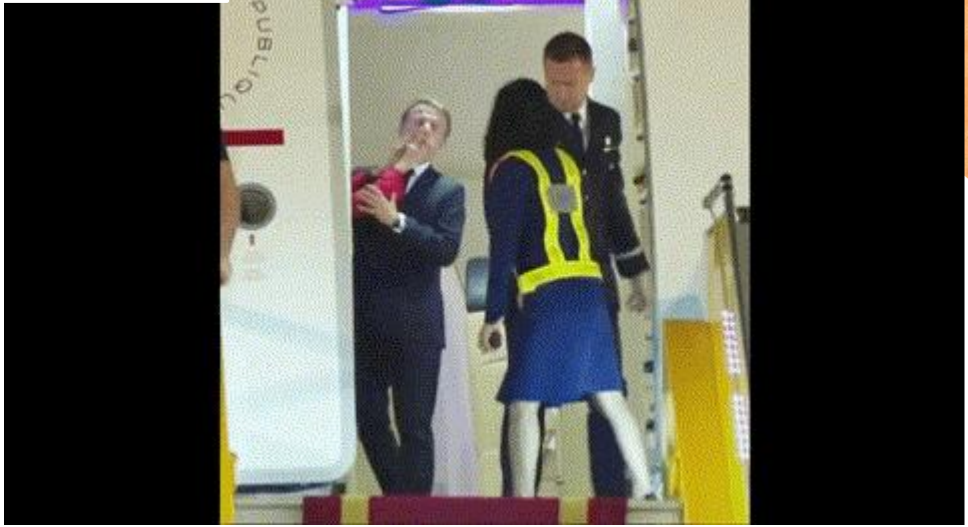
12: Social Media Platforms

https://www.dailymotion.com/playlist/x8e4t4

Google Veo 3 & Google Flow

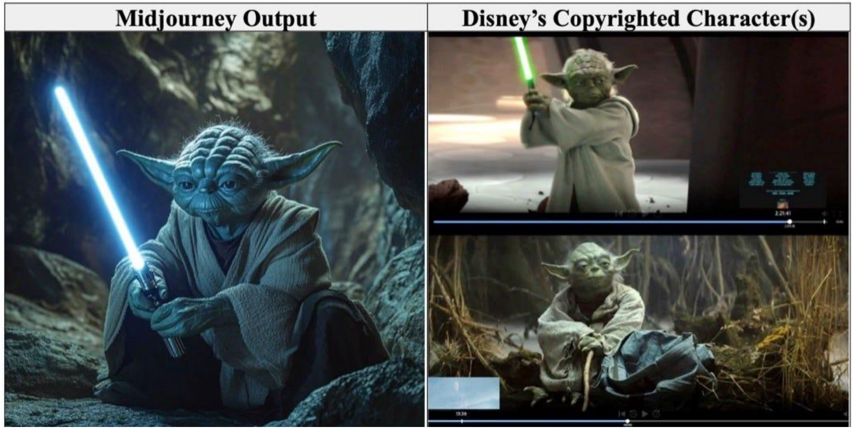


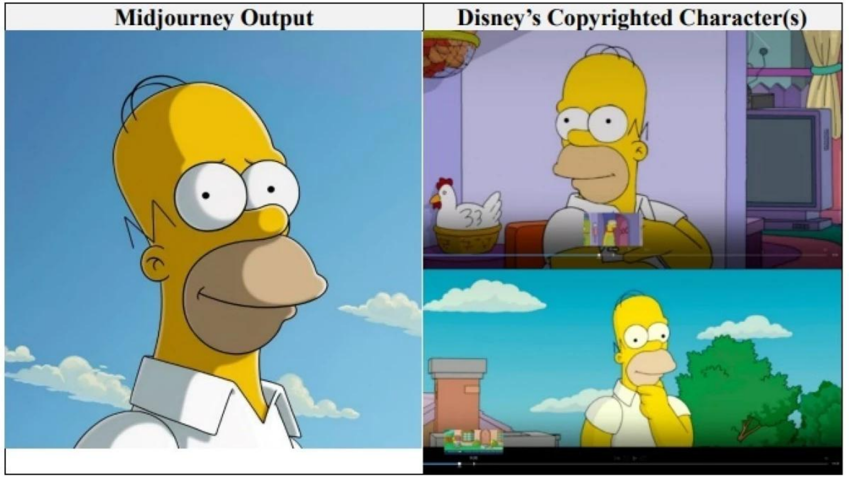
"...swatting a fly"



"...being restrained"

MidJourney Video & Lawsuits!?

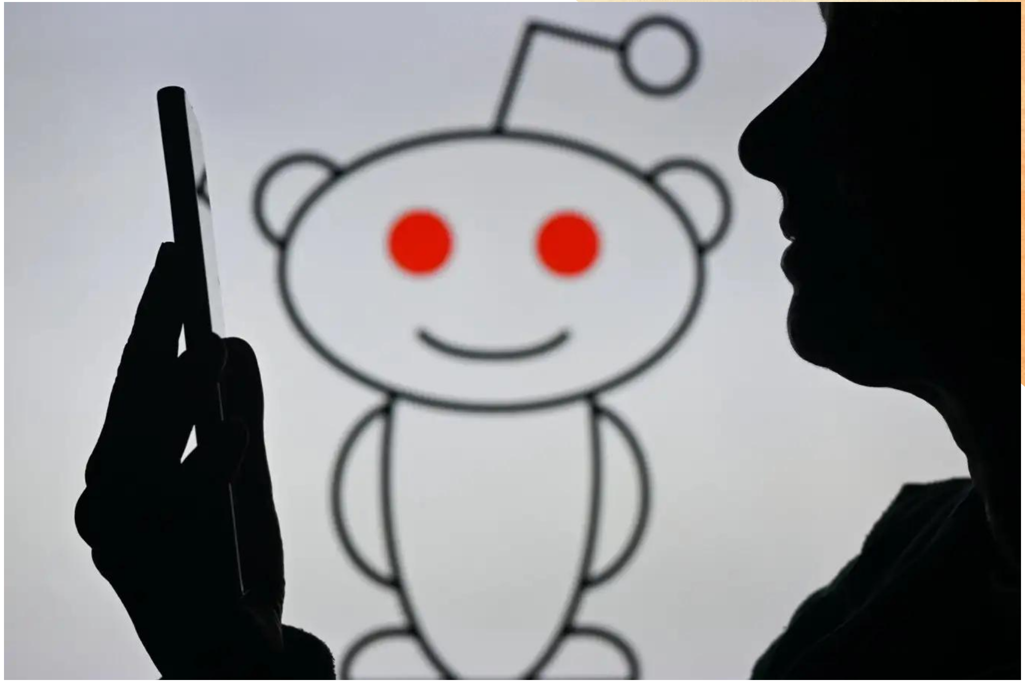




AI Changing Your Opinion?

University of Zurich AI Study: 2025

- AI-generated comments were **six times more persuasive** than human-written responses
    - A University of Zurich AI study on Reddit's r/changemyview subreddit, using deceptive AI bots, caused controversy over ethical concerns and lack of consent, despite demonstrating AI's persuasive potential.



South Korea Digital Sex Crimes

226.9%

increase in digital sex crime reporting at The Digital Sex Crime Victim Support Centre

HYBE & DeepFake Arrests

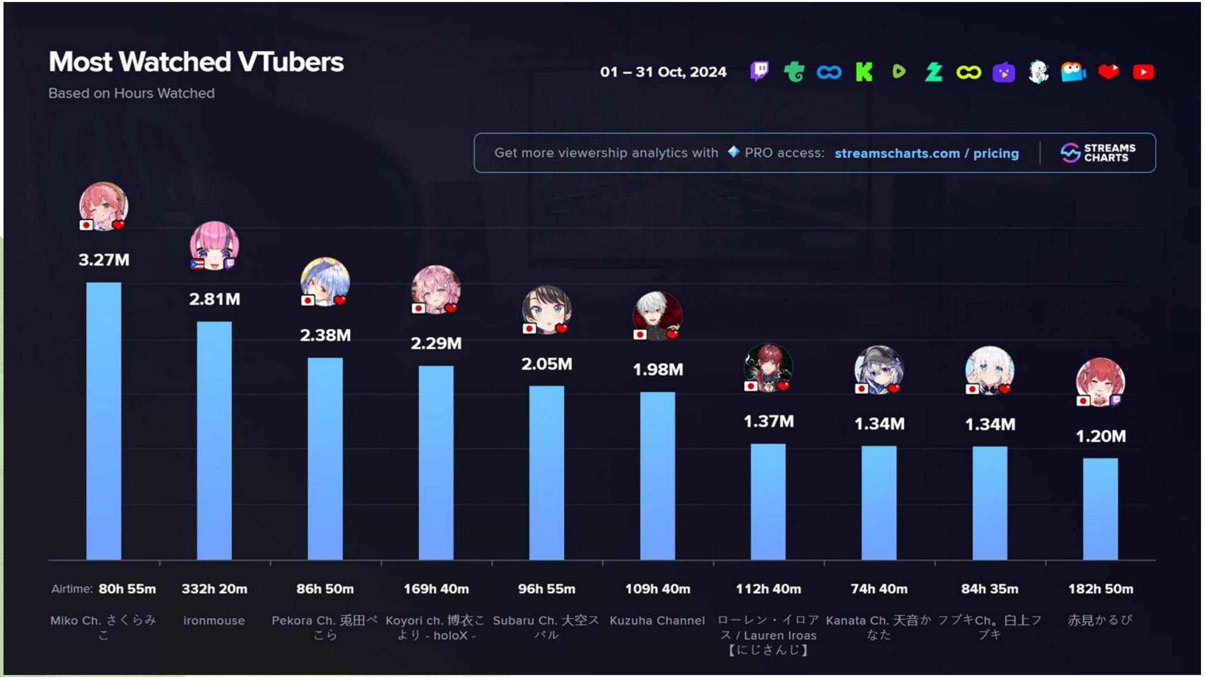
Apr 12. 2025

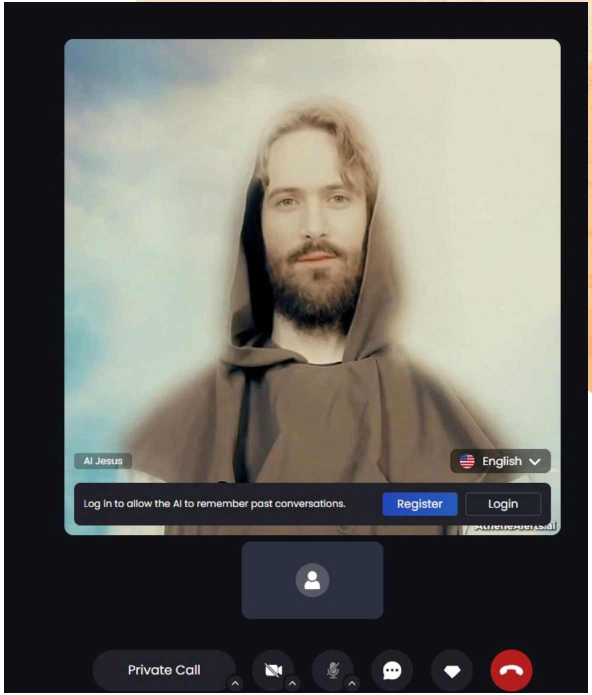
8 arrested in the last week with another **23 under investigations** in Singapore.

HYBE

WE BELIEVE IN MUSIC

2025 AI & Trend Alerts:





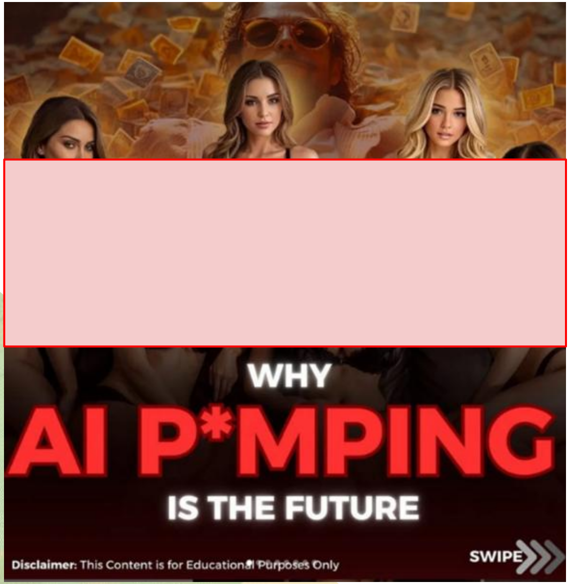
Log in to allow the AI to remember past conversations.

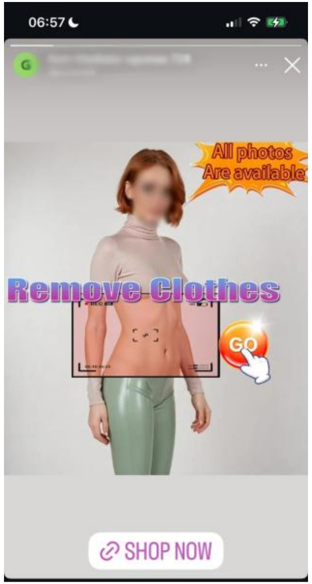
https://athenegpt.ai/video-call?ai=jesus

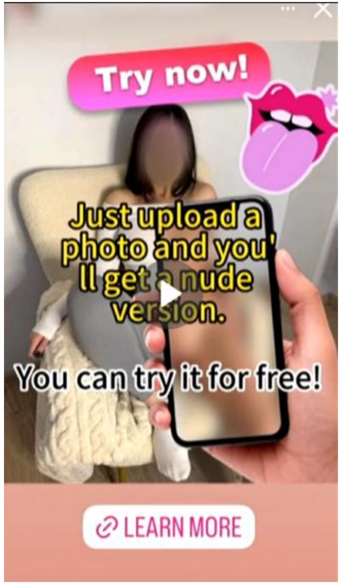
https://www.twitch.tv/ask_jesus

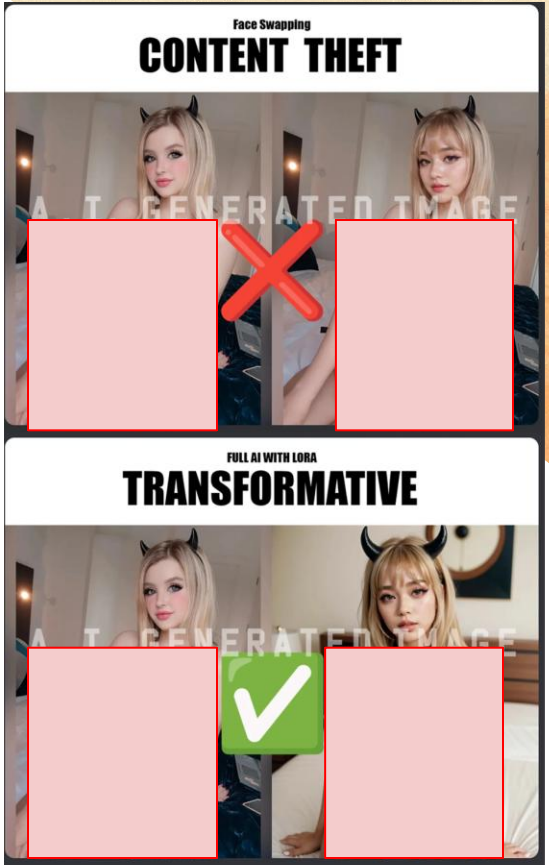
2025 AI & Trend Alerts:

2025 AI “Romance” Updates


[redacted]



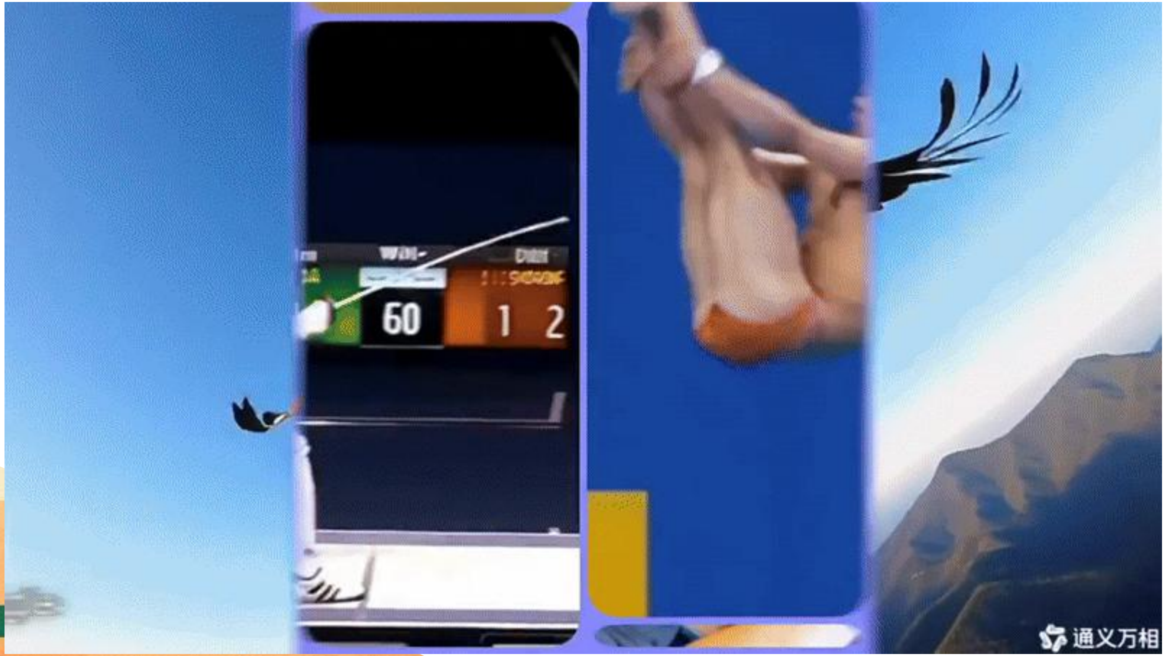



[redacted]
[redacted]
[redacted]
[redacted]

https://www.wired.com/story/ai-pimping-industry-deepfakes-instagram/

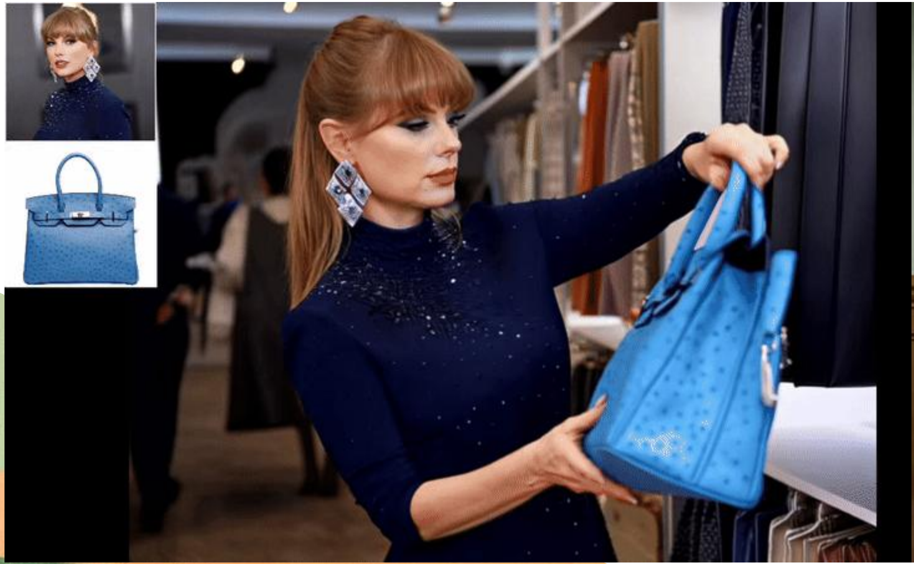
Nation State Actor - AI

Alibaba Cloud, releases its new image generation tool called **WANX**, which has been retroactively used to generate high volumes of sexually explicit content.



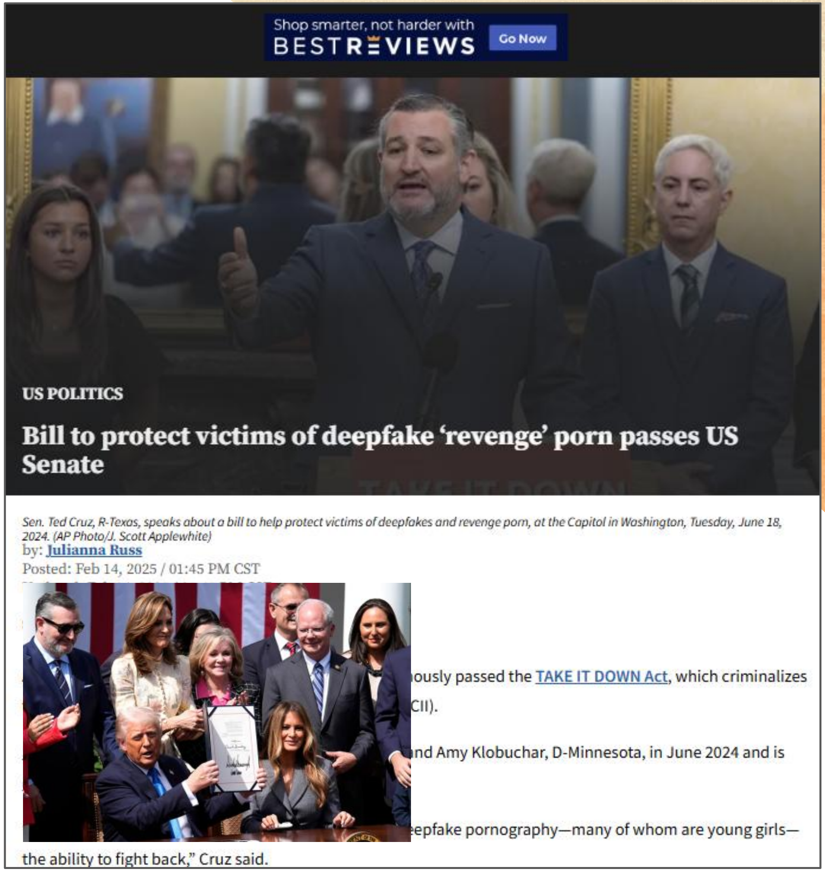
Nation State Actor - AI

Bytedance, releases its new character consistent generation tool called **Phantom**, and their latest open source Google Veo alternative **SeeDance**



2025 AI Updates

LSP: Convicted sex offender among 2 men arrested after investigation into unlawful deepfakes



US POLITICS

Bill to protect victims of deepfake ‘revenge’ porn passes US Senate

Sen. Ted Cruz, R-Texas, speaks about a bill to help protect victims of deepfakes and revenge porn, at the Capitol in Washington, Tuesday, June 18, 2024. (AP Photo/J. Scott Applewhite)

by: Julianna Russ

Posted: Feb 14, 2025 / 01:45 PM CST

ously passed the TAKE IT DOWN Act, which criminalizes

CII).

nd Amy Klobuchar, D-Minnesota, in June 2024 and is

epfake pornography—many of whom are young girls—

the ability to fight back,” Cruz said.

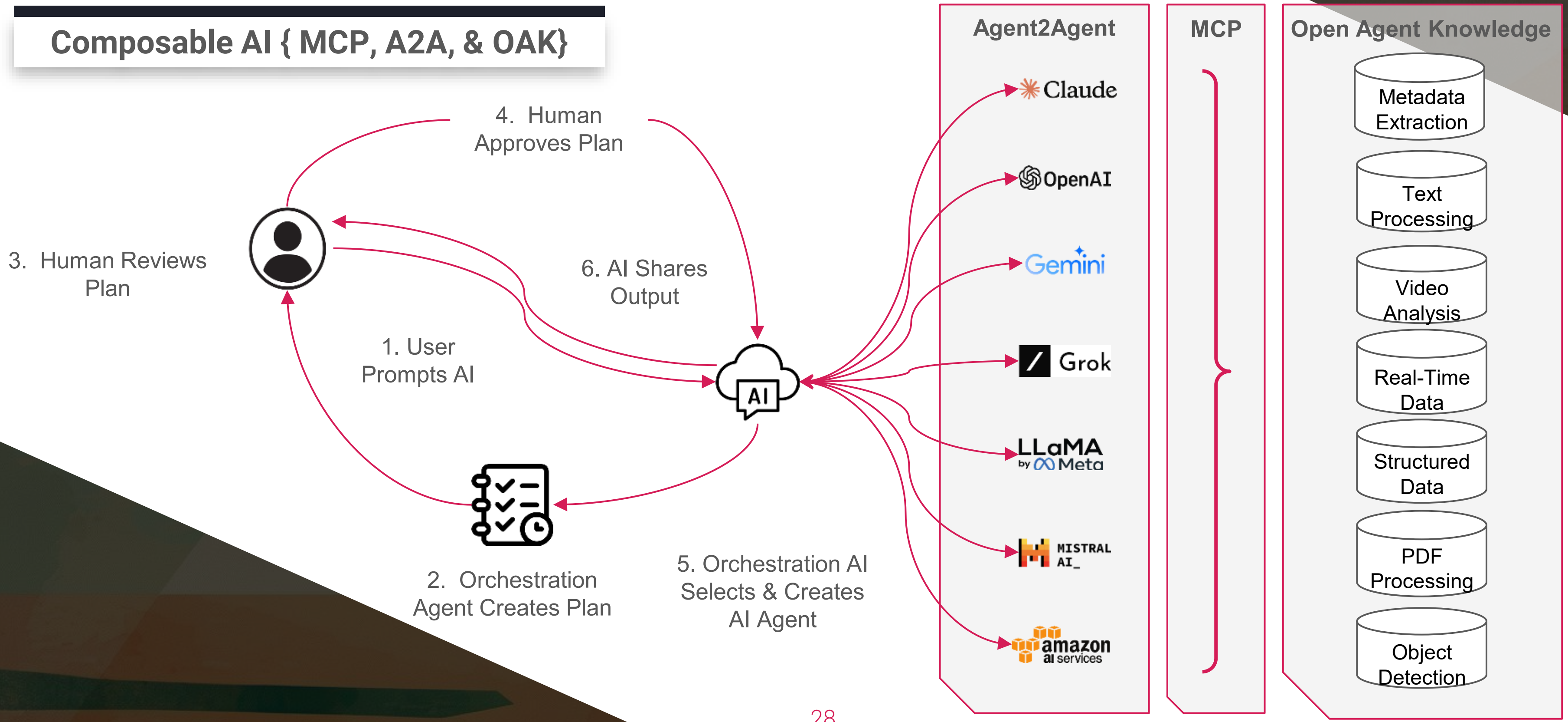
AI Offensive Usage by Private Organizations

Headline:

- O2 has today unveiled the newest member of its fraud prevention team, ‘Daisy’. As ‘Head of Scammer Relations’, this state-of-the-art AI Granny’s mission is to talk with fraudsters and waste as much of their time as possible

AI Systems | 2025 Future of Work

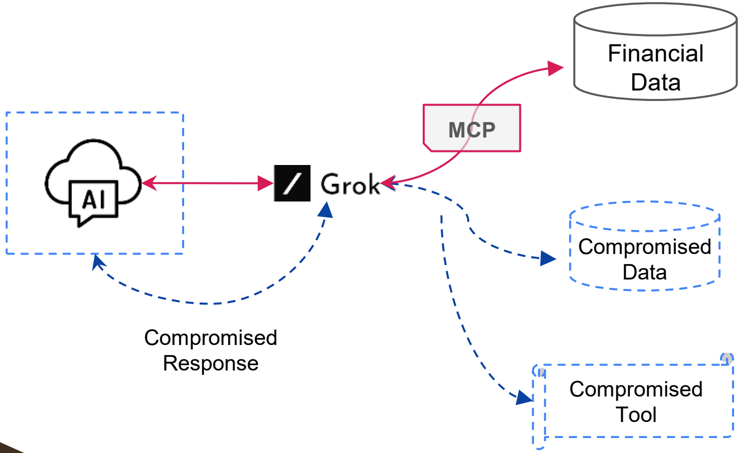
Composable AI { MCP, A2A, & OAK}



MCP | Security Mitigation

- Tool Poisoning: Attackers encode hidden instructions within a tool's name, description, or parameters. These instructions redirect intent, modifying behavior, or injecting subtle hallucinations.
- Malicious External Resources: Even when the MCP server itself is clean, it may route requests through an untrusted third-party API. The external resource then injects prompt-based attacks in returned content, which the LLM consumes as legitimate.

- Agent Planning Isolation: Separate tool ingestion from task planning. Limit context exposure during the planning phase to trusted descriptors only.
- Metadata Sanitization: Implement NLP-based filtering of all incoming tool metadata before ingestion. Red flag patterns resembling system instructions, jailbreak chains, or recursive logic.



You’re Welcome?

Rob Petrosino
Strategic Engagement Advisor - AI | Office of the Private Sector
Rob@vissi.io