AI AND DEEPFAKES

ABA-NCBP ~ February 3, 2024 ~ Louisville KY

Lucy L. Thomson
Chair, ABA Task Force on Law and AI
Livingston PLLC, Washington, D.C.

National Security Agency | Federal Bureau of Investigation | Cybersecurity and Infrastructure Security Agency

Cybersecurity Information Sheet

TLP:CLEAR

Contextualizing Deepfake Threats to Organizations

Deepfakes Pose a National Security Threat

Utilizes AI/ML to create believable and highly realistic media.
Most substantial threats from the abuse of synthetic media: techniques that –

- Threaten an organization's brand,
- Impersonate leaders and financial officers, and
- Use fraudulent communications to enable access to an organization's networks, communications, and sensitive information.

FBI: Synthetic Content

The broad spectrum of generated or manipulated digital content, which includes images, video, audio, and text.

These techniques are known popularly as deepfakes or GANs (generative adversarial networks).

https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats

Public Service Announcement

FEDERAL BUREAU OF INVESTIGATION

June 5, 2023

Alert Number
I-060523-PSA

Questions regarding this PSA should be directed to your local **FBI Field Office**.

Local Field Office Locations:
www.fbi.gov/contact-us/field-offices

Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes

The FBI is warning the public of malicious actors creating synthetic content (commonly referred to as "deepfakes"[a]) by manipulating benign photographs or videos to target victims. Technology advancements are continuously improving the quality, customizability, and accessibility of artificial intelligence (AI)-enabled content creation. The FBI continues to receive reports from victims, including minor children and non-consenting adults, whose photos or videos were altered into explicit content. The photos or videos are then publicly circulated on social media or pornographic websites, for the purpose of harassing victims or sextortion schemes.

Emerging Cyber Attack Vector

IMPACT ON LAWYERS AND THEIR CLIENTS

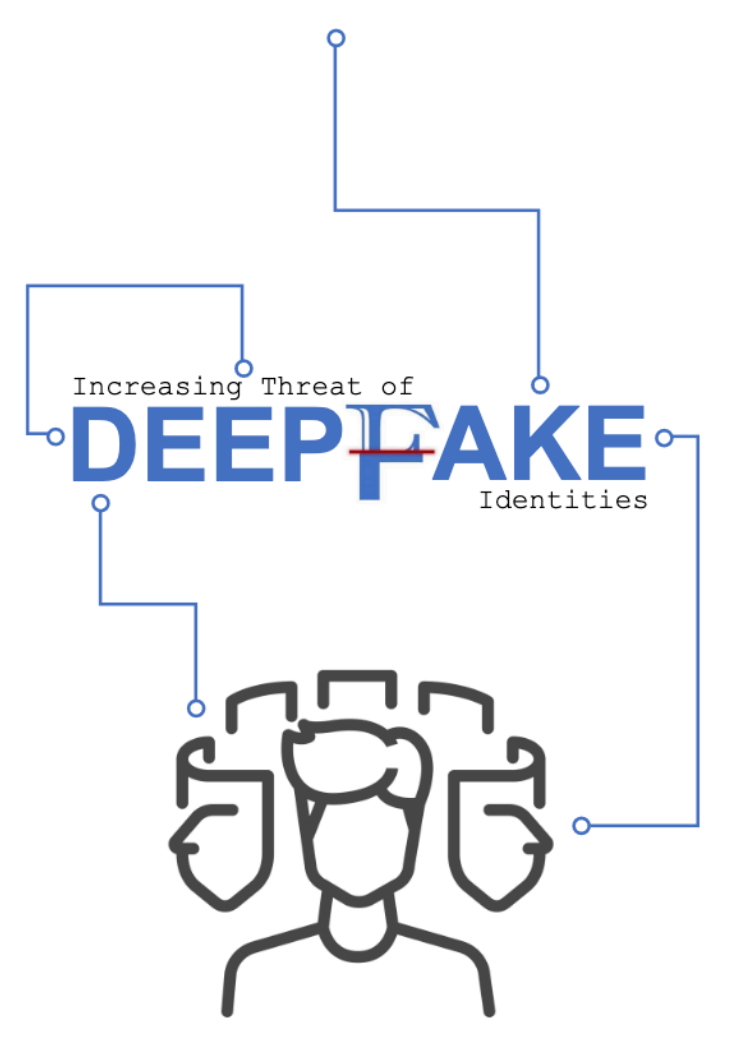
Examples of Inauthentic Content

- Manipulated Audio/Video
- Forgeries
- Proxy/Fake Websites

Business Identity Compromise (BIC)

- Involves the use of content generation and manipulation tools to develop synthetic corporate personas or to create a sophisticated emulation of an existing employee.
- Very significant financial and reputational impacts to victim businesses and organizations.
- Additional harm could be significant, as it would not be difficult to create a deepfake of an emergency alert warning that an attack was imminent or disrupt a close election by releasing a fake video or audio recording of one of the candidates days before voting began.

Commerce Scenario 3: Financial Institution Social Engineering Attack



https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

- In this scenario, the malign actor decides to employ a deepfake audio to attack a financial institution for financial gain.
- Next, she conducts research on the dark web and obtain names, addresses, social security numbers, and bank account numbers of several individuals.
- The malign actor identifies the individuals' TikTok and Instagram social media profiles. She utilizes the videos posted on social media platforms to train the model and creates deepfake audio of targets.
- The malign actor researches the financial institution for the verification policy and determines there's a voice authentication system. Next, she calls the financial institution and passes voice authentication.
- Next, she calls the financial institution and passes voice authentication.
- She is routed to a Representative and then utilizes the customer proprietary information obtained via the dark web. The malign actor tells the Representative that she was unable to access on her account online and needs to reset her password.
- She was provided a temporary password to access the online account. The malign actor gains access to her target's financial accounts.
- The malign actor wires funds from the target's account to overseas accounts.

Homeland Security

Regulating AI Deepfakes and Synthetic Media in the Political Arena

- Policymakers must prevent manipulated media from being used to undermine elections and disenfranchise voters.

  Brennan Center for Justice

  (December 5, 2023)

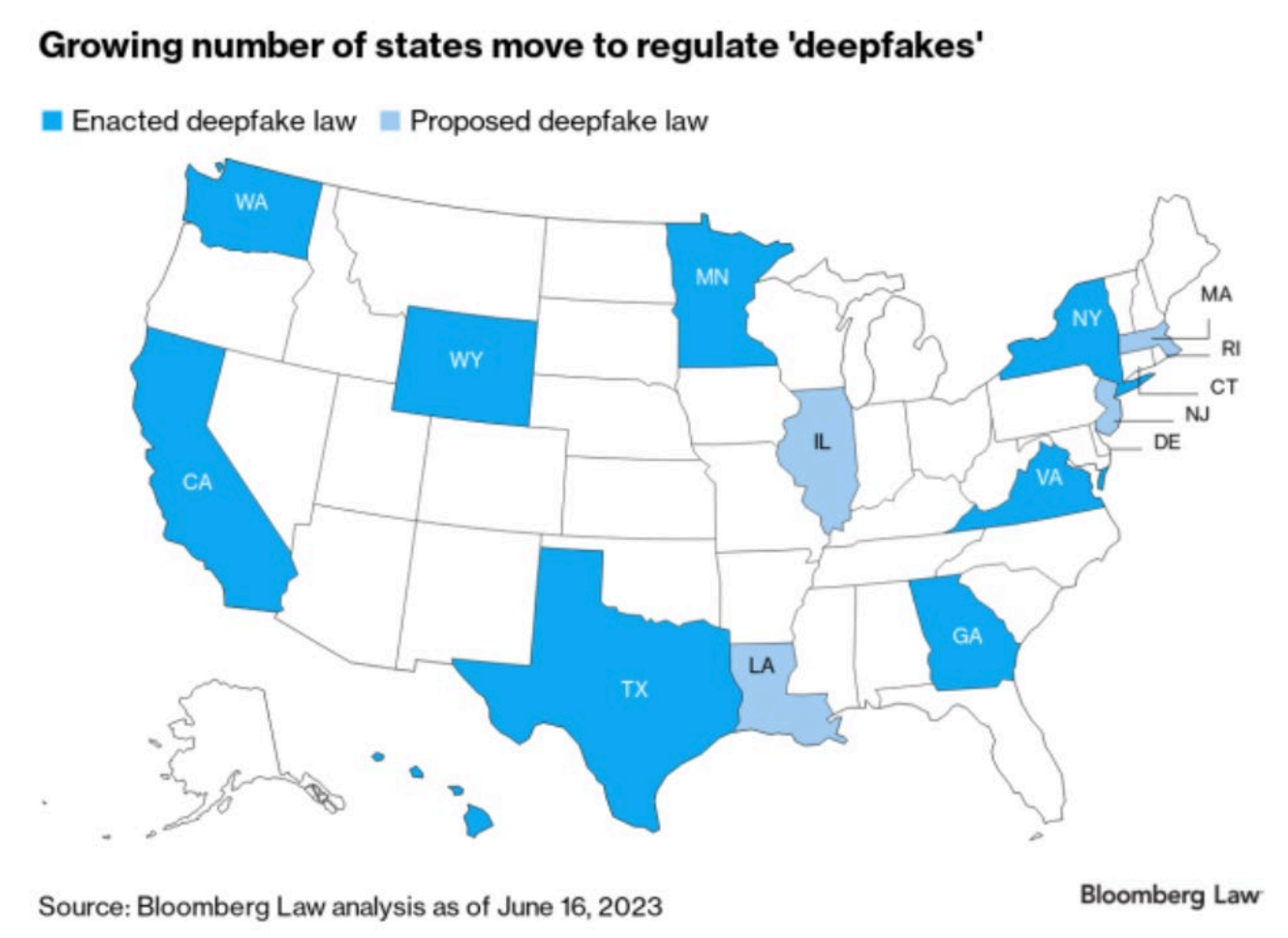
https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena

REGULATING DEEPFAKES

DEEPFAKES Accountability Act of 2023

- Requires creators to digitally watermark deepfake content.
- Makes it a crime to fail to identify malicious deepfakes, including deepfakes depicting sexual content, related to criminal conduct, used to incite violence, and related to foreign interference in an election.
- New provisions would establish an information sharing program to prevent the spread of malicious deepfakes, and impose obligations on social media platforms to include technical capabilities to host content credentials, and requirements around deepfake detection on social media.

Nine states have enacted laws regulating deepfakes — audio and visual forgeries created with AI

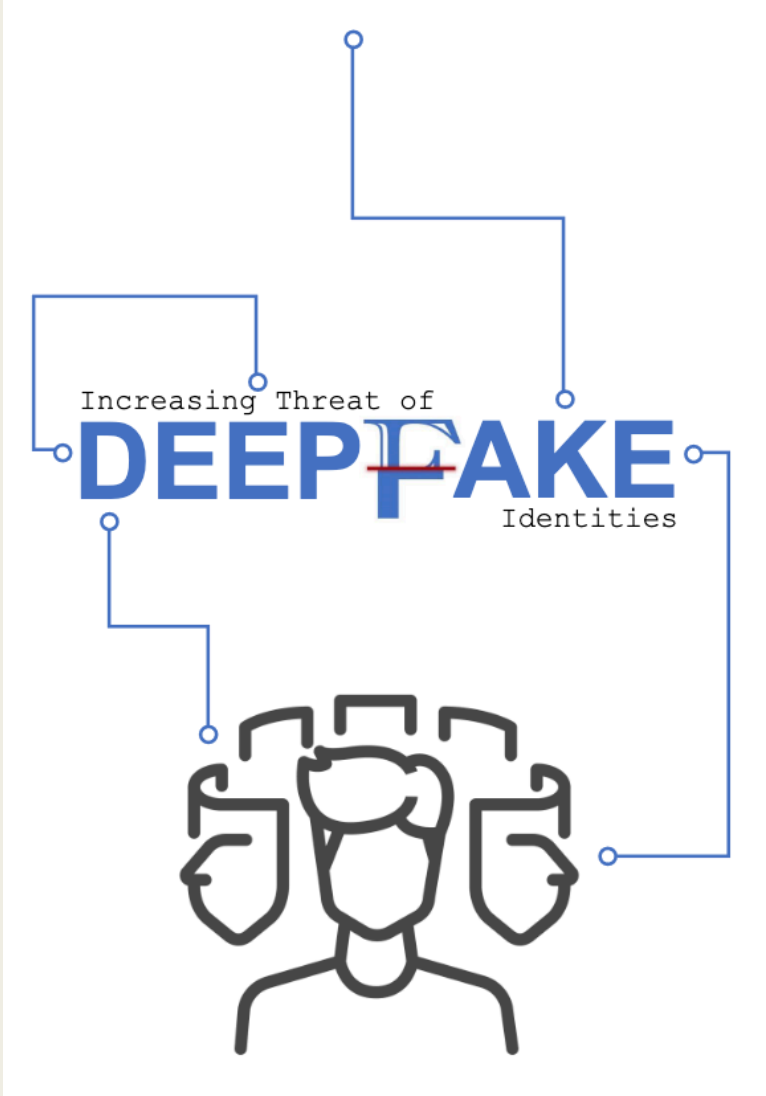


- California, Texas, Washington and Minnesota passed bills that prohibit using deepfakes to influence elections without clear disclosures; Minnesota's law also prohibits using deepfakes to make sexual images of a person without their consent.
- AB-972, CA 2022; SB-751, TX 2019; SB-5152, WA 2023; HF-1370, MN 2023
- New York bill (S1042A) makes it illegal to disseminate AI-generated explicit images or "deepfakes" of a person without their consent. Deepfake porn involves creating fake sexually explicit media using someone's likeness.

AI AND THE COURTS

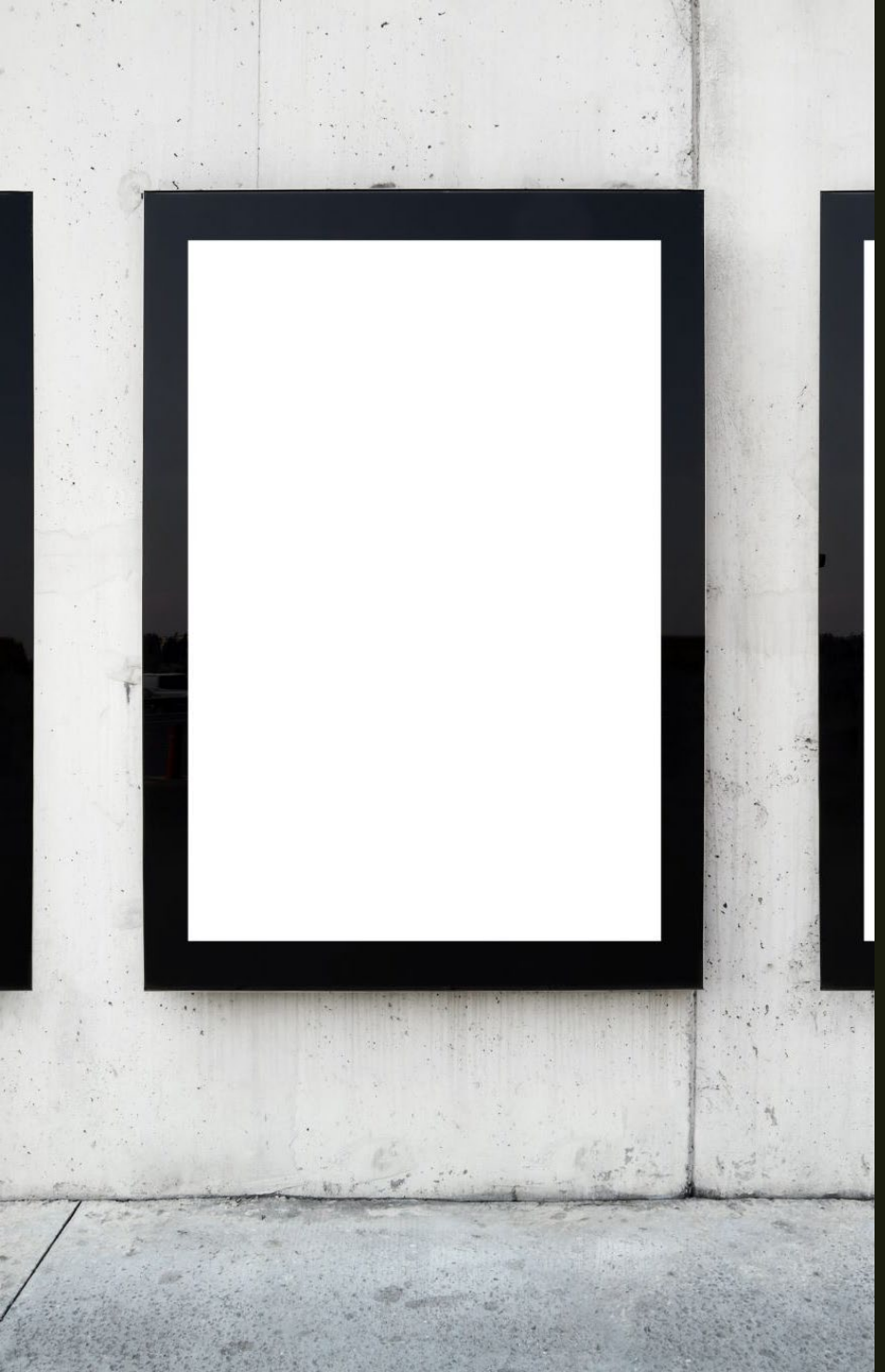
National Security & Law Enforcement Scenario 4: Producing False Evidence in a Criminal Case

- In this scenario, a wealthy criminal defendant, who is accused of murder in a building he owns based on a number of pieces of evidence including latent fingerprints, hair DNA and motive, has objected to the use of identity verification from videos captured in the building lobby based on low resolution and lack of clarity on the face image. As an alibi, he is submitting to the court video imagery from another location in the building that irrefutably puts him elsewhere at the time the crime took place.
- The goal in this case is to weaken the biometric evidence and offer contradictory proof that the defendant has a clear alibi. The deepfake content here is the submitted video itself.
- It is not only the timestamp that can be modified to provide an alibi, but unique circumstances can be created in the video so that it looks genuine. The deepfake video, indirectly, also could also render otherwise strong evidence (such as biometrics) circumstantial, due to the specifics of the building and the scene.



https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

Homeland Security



Can Jurors Spot a Deepfake?

- ***Proposed Modification of Federal Rule of Evidence to address authentication issues regarding artificial intelligence evidence***

A proposed amendment to Rule 901(b)(9) would provide as follows:

(9) Evidence about a Process or System. For an item generated by a process or system:

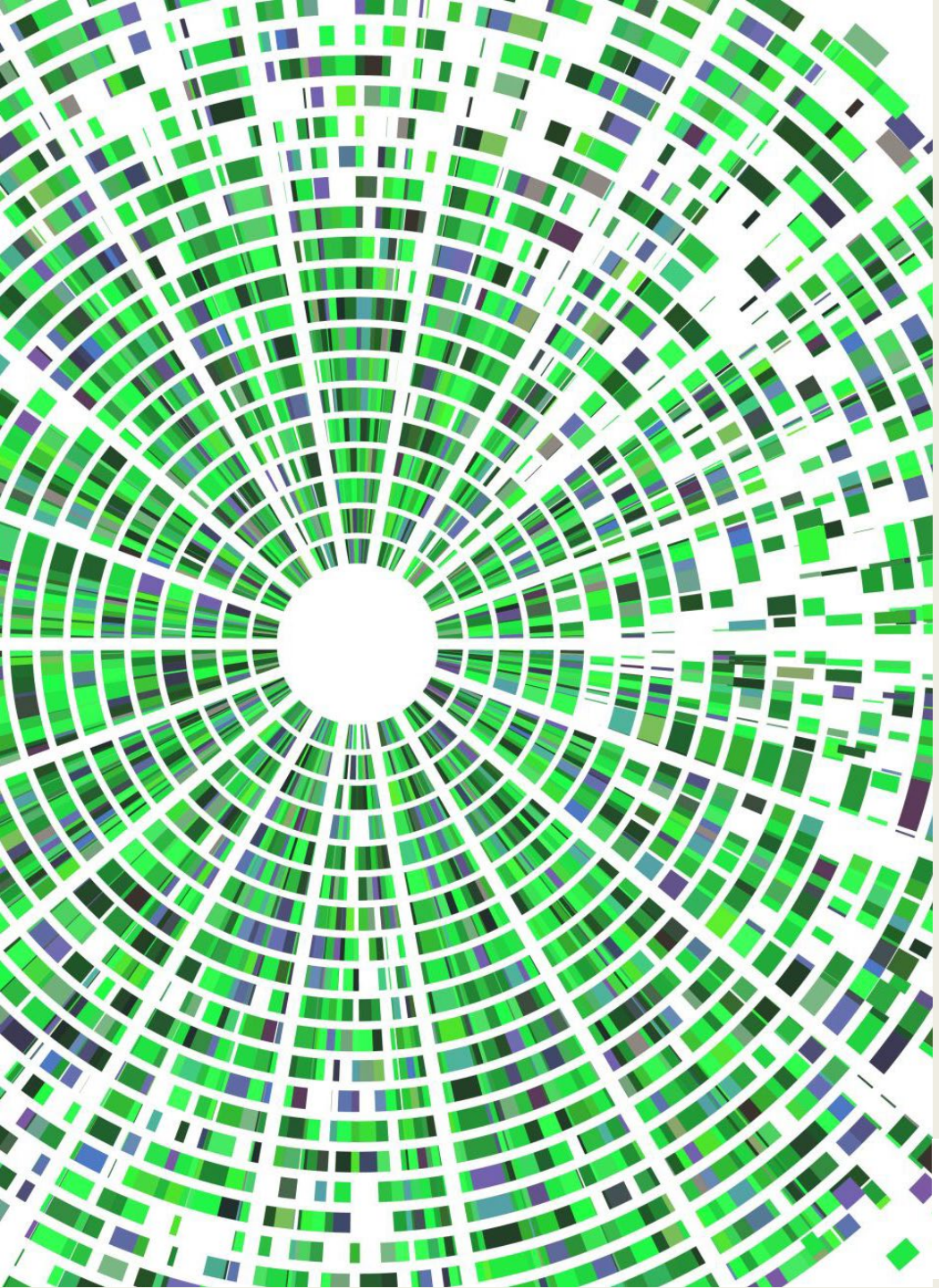
(A) evidence describing it and showing that it produces a *reliable* result; and

(B) if the proponent concedes that — or the proponent provides a factual basis for suspecting that — the item was generated by artificial intelligence, additional evidence that:

(i) describes the software or program that was used; and

(ii) shows that it produced *reliable* results in this instance.

[Hon. Paul Grimm (Ret.) and Dr. Maura *Grossman*



COMBATTING DEEPFAKES

Homeland Security

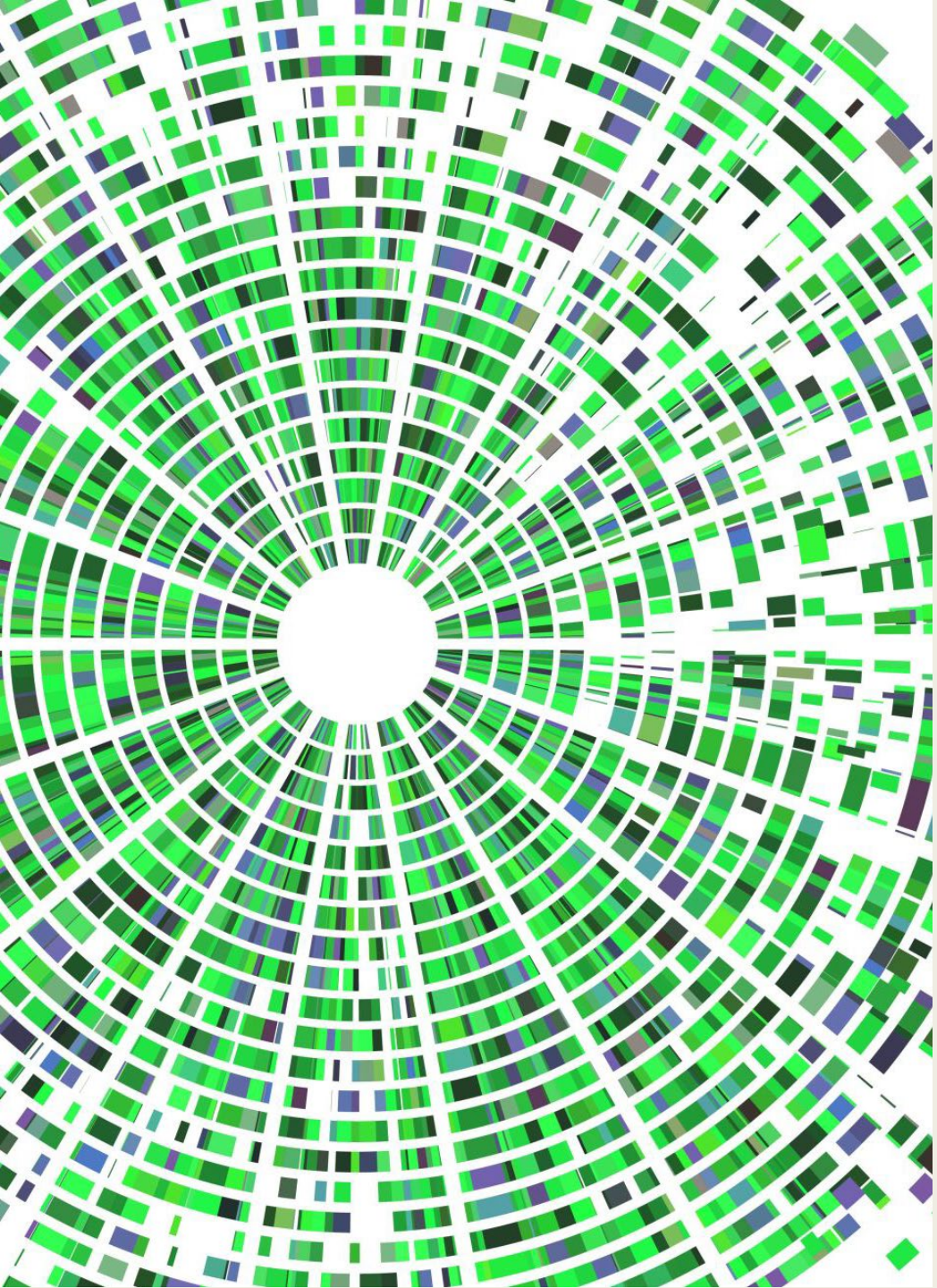
FBI RECOMMENDATIONS FOR RESISTING DEEPFAKES

(1) **Select and implement technologies** to detect deepfakes and demonstrate media provenance, including:

- Real-time verification capabilities
- Passive detection techniques
- Protection of high priority officers and their communications.

(2) **Protect public data of high-priority individuals.**

- Use active authentication techniques such as watermarks and/or CAI standards.
- Minimize the impact of deepfakes:
  - Plan and rehearse: Ensure plans are in place to respond to a variety of deepfake techniques
  - Train personnel



Deepfakes Training Resources

☐ SANS Institute: “Learn a New Survival Skill: Spotting Deepfakes”

☐ MIT Media Lab: “Detect DeepFakes: How to counteract information created by AI” and MIT Media Literacy

☐ Microsoft – “Spot the Deepfake.”

Leveraging cross-industry partnerships

C2PA
Coalition for Content Provenance and Authenticity

Founded by Microsoft and Adobe, includes Arm, BBC, Intel, and Truepic.

Content Authenticity Initiative

A community of media and tech companies, NGOs, academics, and others working to promote adoption of an open industry standard for content authenticity and provenance.

Project Origin
Protecting Trusted Media

Establish a Foundation for Trust in Media
"Misinformation is a growing threat to the integrity of the information eco-system. Having a provable source of origin for media, and knowing that it has not been tampered with en-route, will help to maintain confidence in news from trusted providers."