

Guardians of Forensic Evidence

— Establishing an Evaluation Framework & Guideline for Combating Deepfake Media

James Horan

Manager, Multimedia Language Technologies Group (MLT)

Technology Test and Evaluation Division (TTED)
Information Technology Laboratory (ITL)
National Institute of Standards and Technology (NIST)

WHITE HOUSE: WINNING THE RACE AMERICA'S AI ACTION PLAN – JULY 2025

Combat Synthetic Media in the Legal System

One risk of AI that has become apparent to many Americans is malicious deepfakes, whether they be audio recordings, videos, or photos. While President Trump has already signed the TAKE IT DOWN Act, which was championed by First Lady Melania Trump and intended to protect against sexually explicit, non-consensual deepfakes, additional action is needed.¹⁹ In particular, AI-generated media may present novel challenges to the legal system. For example, fake evidence could be used to attempt to deny justice to both plaintiffs and

¹⁹ TAKE IT DOWN Act, Pub. L. No. 119-12, 139 Stat. 55 (2025) (codified as 47 U.S.C. § 223(h)).

AMERICA'S AI ACTION PLAN

defendants. The Administration must give the courts and law enforcement the tools they need to overcome these new challenges.

Recommended Policy Actions

- Led by NIST at DOC, consider developing NIST's *Guardians of Forensic Evidence* deepfake evaluation program into a formal guideline and a companion voluntary forensic benchmark.²⁰

A LONG HISTORY OF MULTIMEDIA RESEARCH AT NIST

NIST has done extensive research and has a long history of evaluation of media analytics, provenance, authentication and manipulation

Project	Category	Official Resource Link
MediFor / SemiFor	Forensics	arpa.mil/program/media-forensics arpa.mil/program/semantic-forensics
OpenMFC	Forensics	mfc.nist.gov
TRECVID	Video Analytics	trecvid.nist.gov
DIVA / ActEV	Video Analytics	actev.nist.gov
Aladdin	Video Search	arpa.gov/research-programs/aladdin-video
SRE	Speech Analytics	sre.nist.gov
ARTS / Speech	Speech Analytics	nist.gov/programs-projects/speech-analytics
PSIAP / ASAPS	Public Safety	nist.gov/ctl/pscr/research-portfolios/archived-portfolios/analytics
CCU	Multimodal Analytics Social & Behavioral AI Evaluation	arpa.mil/program/computational-cultural-understanding

PROBLEM: SHIFT FROM "VIDEO DOESN'T LIE" TO UNIVERSAL SKEPTICISM

The Forensic Paradigm Shift: The End of Implicit Trust

- **Old Standard:** Chain of Custody (Is this the original file?)
- **New Standard:** Content Authentication (Is the reality in the file genuine?)

Note: Law enforcement can no longer assume video/audio is genuine.

Resource Strain & The Forensic Gap: The Reality of Modern Investigations:

- **Past:** Authentication took ~30 minutes.
- **Present:** Authentication requires days of expert analysis and specialized AI detection software.

Need: Immediate investment in digital forensic training and tools.

THE CRITICAL CHALLENGE: DEEPFAKE MANIPULATION

The Problem: EROSION OF TRUST

The surge of deepfakes and AI-generated media has sparked a crisis of authenticity that is quickly outstripping current legal frameworks. As generative tech evolves, traditional forensic methods are struggling to keep pace with these increasingly sophisticated adversaries.

The Risk: Miscarriages of justice through fabricated digital evidence

- ✓ High-Fidelity Visual Forgery: Deepfake image and video manipulations that bypass traditional forensic detection and deceive human perception.
- ✓ Audio voice cloning & tampering: : The use of generative AI to create near-perfect vocal imitations or alter existing recordings to misrepresent testimony.
- ✓ Automated Disinformation Campaigns: The rapid deployment of synthetic media at scale to manipulate public opinion or contaminate legal discovery processes.



Image credit: Google Gemini (Nano Banana)

AI DEEPFAKE FRAUD: MAJOR RECENT INCIDENTS



Case Name	Target / Victim	Fraud Method	Loss / Impact
Wrong Evidence (Melissa Sims Case) ¹	Florida Woman	Ex-boyfriend used AI to forge text messages; police failed to verify them before making an arrest.	Innocent Jail Time (2 days); exposed massive gaps in digital evidence verification.
"Liar's Dividend" (Legal Threat) ²	Courts & Juries	AI-generated deepfake video submitted as evidence	Case dismissed; risk of wrongful rulings; erosion of trust in legal evidence
The Virtual Heist (Arup) ³⁻²	Arup (British Company)	Real-time deepfake video call impersonating the CFO and multiple colleagues.	\$25.6 Million in stolen corporate funds.
The "Musk" Trap ³	Social Media Users	Viral AI-cloned video ads of Elon Musk promoting fake investment platforms.	\$690,000+ (single victim); global scale of fraud.
The School Frame ³⁻⁴	The Pikesville High School Principal	Malicious AI audio clip of fake "racist remarks" created by an internal rival.	Reputational ruin and death threats; suspect arrested.
The CEO Ruse ³⁻⁵	WPP Advertising	AI voice clone used during a live Teams meeting to impersonate the CEO.	Thwarted by alert staff; highlighted new real-time voice attack vectors.

1. <https://www.youtube.com/watch?v=594t2W2961o>

2. <https://www.youtube.com/watch?v=kcbLvCfH7Lo>

3. <https://www.incode.com/blog/top-5-cases-of-ai-deepfake-fraud-from-2024-exposed>

CRITICAL THREATS TO JUSTICE

- ✓ **Fabricated Alibis:** Suspects using AI to place themselves away from a crime scene (e.g., fake gas station footage).
- ✓ **Character Attacks:** Fake Body-Worn Camera footage showing officers planting evidence or using slurs to incite unrest.
- ✓ **Witness Confusion:** High-quality deepfakes on social media overwriting a witness's actual memory of an event.
- ✓ **The "Liar's Dividend":** Exploiting Public Skepticism: Defendants falsely claim that legitimate, incriminating footage is "AI-generated" to manufacture reasonable doubt and undermine the truth.



THE PROGRAM SCOPE

Empowering the Forensic Examiner with Guideline & Evaluation Framework

- ✔ Strategic Positioning: We are not developing a new detector. We are establishing an Evaluation Framework to stress-test SOTA algorithms against forensic realities.

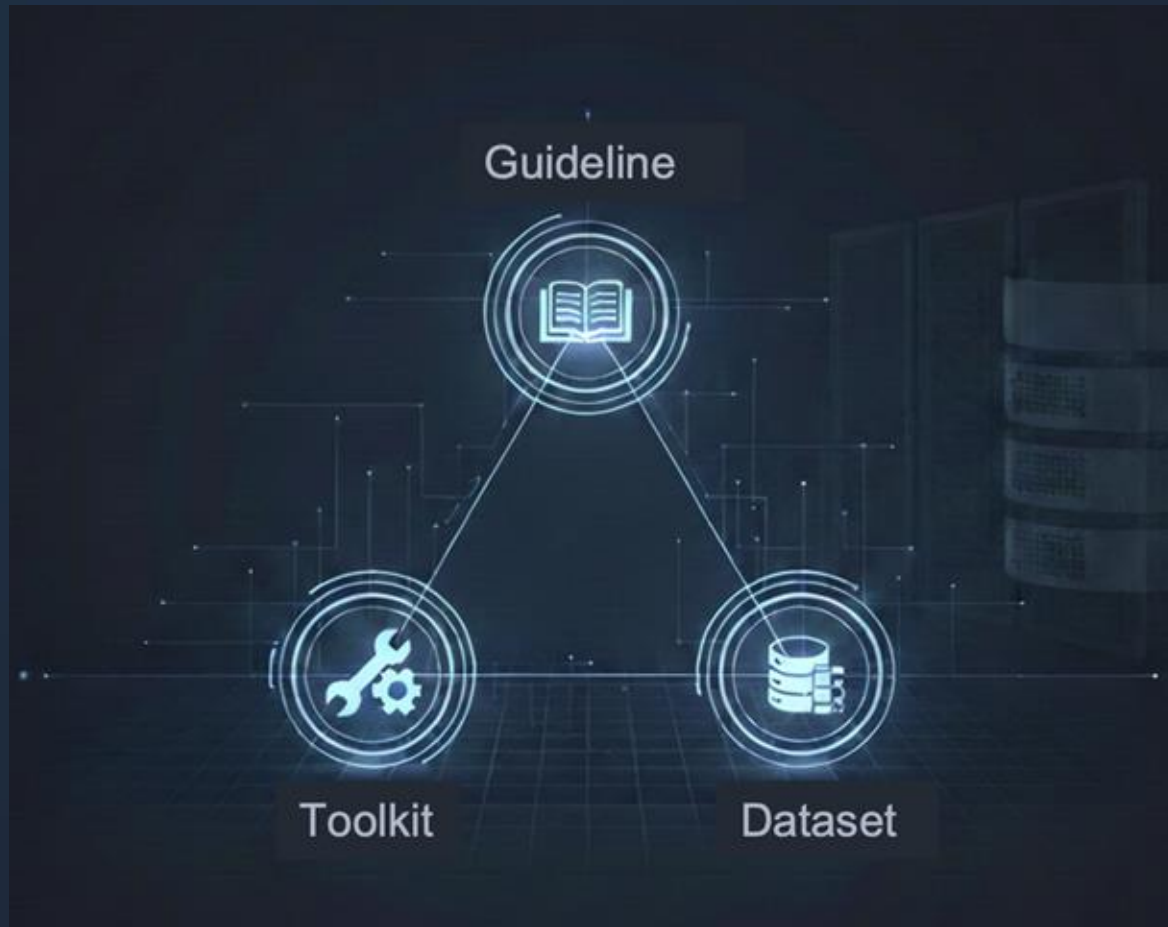
Bridging the Gap: From Lab Accuracy to Forensic Practice

- ✔ The Forensic Reality Check: High-performing academic detectors (90%+ accuracy) often fail when exposed to "Forensic Realities."
Stress-Test Factors: Media compression, low-bitrate surveillance footage, social media laundering, multi-generational re-encoding, and anti-forensics.

Raising Awareness of Algorithmic Limitations in Legal Contexts

- ✔ The Call to Action: Both forensic and academic communities must collaborate to ensure research aligns with the actual needs of the practitioner.

The SOLUTION



Forensic Guidelines

- ✔ Providing examiners with a comprehensive framework for digital investigations .

Deepfake Challenge Toolkit

- ✔ Delivering a specialized toolkit designed for the evaluation of deepfake analytic tools.

Forensic Deepfake Dataset

- ✔ Providing the forensic community with a comprehensive dataset for research and validation.

CALL FOR COLLABORATION

BRIDGING RESEARCH & PRACTICE

- ✓ We are merging academic innovation with the absolute rigor of forensic science.
- ✓ By collaborating, we ensure that digital evidence presented in the courtroom remains secure, defensible, and trusted.

JOINING THE WORKING GROUP

GUIDELINES

ACTIVITIES AND NEXT STEPS

- ❑ NIST successfully submitted the 120-day deliverable Project Plan for a multiple year program
- ❑ Started outreach and engagement process – Starting outreach for a forensic expert group: USSS, FBI, state & local LE, academic R&D (UBuffalo, Drexel), forensic process and training experts
- ❑ Working on data options, tool options and framework for forensic practitioners including “DeepFake Challenge Kit”
- ❑ Landscape survey of industry capability
- ❑ Reviewing existing guidelines and best practices in this domain
- ❑ Investigating possible SDO’s

TAKE AWAY

"Trust is the currency of the digital age.
We must secure it with science."

— Guardians of Forensic Evidence (GoFE) Initiative

Q & A

deepfake@nist.gov

REFERENCE SOURCES

Note on Visual Assets: The conceptual images, diagrams, and figures contained within this presentation were generated by Gemini (Nano Banana model). These assets are intended for illustrative purposes to support the conceptual framework of the project.