INTERPOL

FALSE FACADES

Policy Approaches to Deepfake Detection

GLOBAL GUIDELINE

AUGUST 2026

Project funded by the
Government of Japan

TABLE OF CONTENTS

FOREWORD

Emerging synthetic media threats require the law enforcement community to upgrade its detection and mitigation skills. Recognizing the need for understanding the complicated layers of synthetic media production, identification, and deconstruction, INTERPOL offers this guideline to law enforcement practitioners dealing with threats and crimes facilitated through videos, images, and audio created or modified using artificial intelligence.

Synthetic media have the ability to permeate different crime categories due to their ease of production and dissemination. Moreover, there are various degrees with the depth of manipulation that could easily compound the difficulty of identifying different types of synthetic media. Crimes get more sophisticated as threat actors figure out ways to make synthetic media as realistic and natural as possible, whether this means incorporating imperfections or finding ways to be incompatible with automated detection.

Synthetic media are being weaponized in non-consensual deepfake sexual abuse material, financial fraud, disinformation, identity theft, and extortion – reflecting a systemic, transnational threat that spans cybercrime, financial crime, organized crime, and counter-terrorism.

This guideline provides information on the spectrum of synthetic media, from shallow fakes to deepfakes to hybrid formats. Tools and techniques that may be used for authentication, as well as relevant adoption standards and legal considerations, are also illustrated in this document.

As one of the main components of Project SynthWave, an initiative which aims to support INTERPOL member countries in enriching their capabilities in tackling synthetic media threats, this guideline presents crucial discussions on the chain of custody, detection approaches and education, and AI and machine learning.

I express my appreciation to the Government of Japan for supporting Project SynthWave and providing law enforcement agencies a platform to start and maintain important conversations and knowledge sharing on synthetic media, not only within our own sector but also with industry and academia.

Together with member countries against crime, INTERPOL strives to equip law enforcement with technical know-how and additional investigative capabilities. I am optimistic that this guideline will aid member countries in our collective fight against cross-cutting crimes brought about by synthetic media.

安原俊伸

Toshinobu Yasuhira
Director, Innovation Centre
Executive Directorate Capacity Building and Innovation
INTERPOL

ACKNOWLEDGEMENTS

INTERPOL would like to extend its gratitude to the Government of Japan for their generous support for Project SynthWave, under which this guideline was compiled. We would also like to thank the law enforcement agencies of the participating countries under Project SynthWave – Brunei, Cambodia, Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand, and Viet Nam – for their active engagement and vital contributions to the project and its initiatives.

INTERPOL would like to acknowledge Ms Emily Williams Ba (Hons) MSc ACSFS AFHEA, PhD Researcher – Synthetic Media Authentication in Forensic Evidence, Liverpool John Moores University, Mr Brian Sheil Ba(Hons) PGCert MSc ACSFS, Director, Brytestar, Ms Grace Robinson BSc MA(Hons), Postgraduate Researcher – Global Security and International Relations, University of Glasgow, Mr Colin Robinson BSc MCSFS, Programme Leader MSc Audio and Video Forensics**,** Liverpool John Moores University, and Mr Darren Wright Ba(Hons) PGCert MSc MCSFS, Deputy Director of Operations, Fiosrú – Office of the Police Ombudsman, Republic of Ireland, for their invaluable research and extensive contributions in writing this guideline.

We are especially grateful to the aforementioned experts for their input to Chapter 7 titled "*Regulation and Legal Consideration*". Their insights were instrumental in enhancing the guideline's global relevance for law enforcement agencies worldwide.

Additionally, we would like to thank the National Police Agency (NPA) Japan and the Japan Cybercrime Control Center (JC3) for co-hosting the INTERPOL Conference on AI in Digital Forensics: Unpacking Insights in Investigations and Analysis which took place in October 2025 in Tokyo. The conference provided a critical platform for presenting, refining, and advancing the findings contained in this guideline.

Through initiatives such as Project SynthWave, INTERPOL remains committed to empowering law enforcement agencies worldwide with the knowledge, tools, and collaborative frameworks needed to effectively counter the criminal misuse of emerging technologies, including synthetic media.

DISCLAIMER

This guideline offers general information and guidance on understanding and approaching the threats of synthetic media and how law enforcement agencies may respond to them. The information in this guideline is obtained from member countries, private partners, and open sources, and is provided for the consideration of readers at their discretion.

The examples, descriptions, and discussions in this guideline are intended as options for consideration, rather than as recommendations, encouragement, or definitive proposals. References to external links, publications or any actions, proposals, measures, or policies developed on the basis of this guideline must be taken with reference to the applicable laws as verified and tested in the relevant jurisdictions by the relevant readers. INTERPOL bears no responsibility in respect of any such actions, steps, measures, or any documents created based on this research guide.

Links to external publications or websites included in this guideline are provided as references only, and do not constitute an endorsement by INTERPOL of those publications or their content. It is the responsibility of the user to evaluate the content and usefulness of information obtained from such other publications/websites.

Descriptions of the provisions of certain legal instruments in this document are presented as discussions only and are not, nor may be construed as, proposals on applicable interpretations in respect of any of these legal instruments.

This guideline has not been formally edited. The content of this publication does not necessarily reflect the views or policies of INTERPOL, its member countries, its governing bodies, or contributory organizations, nor does it imply any endorsement.

Although this guideline does not contain law enforcement information, it should still be considered sensitive and should be protected from unauthorized access and should only be used for law enforcement purposes. Any further dissemination of this content is restricted to official law enforcement authorities for legitimate law enforcement purposes. This information may be used solely by national law enforcement agencies for information and international police cooperation purposes, may only be used for intelligence purposes, and should not be construed as being a document of evidential value or for use in judicial proceedings. Any decision on action to be taken based on this information is within the discretion and responsibility of the appropriate national law enforcement authorities and should be done in accordance with the applicable national law and international obligations of the country concerned. The written permission of INTERPOL is required in order to disseminate this document or any part of it.

EXECUTIVE SUMMARY

This document addresses the escalating threat of synthetic media, defines their denominations, and advocates for a strategic pivot from mere detection to robust media authentication, upholding public trust and legal integrity. The guideline specifically focuses on audio, video, and image-based media, where manipulation can be highly convincing and often evades traditional detection mechanisms. By introducing the key themes, issues, and solutions that are currently at the forefront of forensics and policing, this guideline aims to equip both practitioners and decision makers with both awareness and understanding of what they are currently facing.

Deepfakes, generated by advanced AI techniques such as generative adversarial networks (GANs) and diffusion models, are becoming increasingly indistinguishable from authentic media. This technological sophistication poses a significant challenge to institutional readiness. Deepfakes are designed to appear "native" and can incorporate contextual imperfections (the Leinster Principle), such as filters or noise, to mimic real-world flaws and bypass both human and AI detection systems. Current deepfake detection tools, while showing promise, often prove inaccurate, inconsistent, and opaque (black box algorithms) in forensic and legal contexts, providing probabilistic and ambiguous outputs that undermine their credibility. Traditional forensic methods like frame analysis and spectral fingerprinting are often overlooked as possible tools for distinguishing the fake and manipulated from the real and original. This document brings some of these techniques to light that will allow practitioners to authenticate for synthetic media, in their own right, or as an added step after AI system identification of suspect files.

This guideline also discusses the use of AI and machine learning as a tool within the forensic workflow, balancing time, cost, efficacy, explainability, and understanding, while upholding the principles of classic forensic science. In addition to this, there are risks and benefits of relying on these tools alone, opening debate on whether this is appropriate within law enforcement.

Identified herein is a necessary shift towards media authentication. Rather than solely taking the outlook of attempting to "catch fakes," the focus is shifted to verifying what is real, establishing trusted chains of custody, validating provenance, and applying consistent forensic standards to confirm authenticity from the outset of an investigation. Provenance, the origin and history of an article of evidence, is highlighted as a critical pillar for prevention, offering a contextual framework, tamper detection points, and legal admissibility – "provenance as detection." Without robust provenance, even authentic media can be rendered unusable, due to doubt raised as a result of the liar's dividend.

Ultimately, the deepfake era necessitates a reconsideration of what counts as "fake", demanding both technical precision and procedural clarity. Requiring new standards, tools, and protocols for media authentication, alongside vital training and education for investigators, forensic analysts, and legal professionals is now non-negotiable.

1. INTRODUCTION

This guideline continues the trajectory of inquiry established in *Beyond Illusions*[1], which first illuminated the emerging risks of synthetic media and the growing influence of AI-generated content on public trust, legal integrity, and national security. Where that foundational work diagnosed the scale of the issue, this guideline aims to operationalize a response, delving deeper into the methods of prevention, detection, and regulation, while identifying the unique forensic challenges posed by modern generative tools. Our emphasis is on audio, video, and image-based media, where manipulation is not only visually or aurally convincing, but can often evade traditional detection mechanisms.

1.1 Deepfakes

At the heart of this issue is a clash between technological capability and institutional readiness. Deepfakes, which are synthetic media generated through advanced machine learning techniques such as generative adversarial networks (GANs) or diffusion models, are increasingly indistinguishable from authentic recordings and/or media. GANs work by having two neural networks: the generator, which creates fake data, and the discriminator, which tries to detect fakes, compete, and improve through repeated iterations. Diffusion models add noise to training data and then learn to reverse this process, reconstructing high-quality, realistic images. In some areas, diffusion models are now outperforming GANs by producing more detailed patterns and textures.

As the line between real and fake continues to blur, the need for a clear framework for classification, analysis, and admissibility of evidence becomes increasingly urgent.

To frame this conversation, we adopt a working definition of AI and synthetic media aligned with that used in the *Beyond Illusions* paper: *"Synthetic media refers to the type of media content, such as images, videos, audio, or text that has been completely or partially generated or manipulated using artificial intelligence (AI) algorithms. These technologies allow for the creation of highly realistic and convincing media that can be difficult to distinguish from human-generated content".*[2]

Within this definition, deepfake content typically originates from machine learning models trained on large datasets of human appearance or voice, enabling the generation of hyper-realistic imitations. These outputs, whether malicious or benign, raise complex questions about authenticity, authorship, and intent.

While much of the current discourse focuses solely on deepfake detection, this guideline proposes a necessary shift in emphasis toward media authentication. Rather than solely attempting to catch fakes, greater value may also lie in verifying what is real, establishing trusted chains of custody, validating provenance, and applying consistent forensic standards to confirm authenticity from the outset.

Deepfake detection tools, while promising in theory, often fall short in forensic or legal settings due to fundamental limitations. Their results are frequently inaccurate, incomplete, and inconsistent, failing to account for traditional manipulations or reliably detect AI-generated content across varied conditions. Many operate as opaque "black boxes", offering little transparency or repeatability, which undermines

[1] INTERPOL (2024). *Beyond Illusions: Unmasking the Threat of Synthetic Media for Law Enforcement*. Available at: https://www.interpol.int/content/download/21179/file/BEYOND%20ILLUSIONS_Report_2024.pdf, last accessed 13 February 2026.
[2] INTERPOL, 2024.

their credibility in court. Additionally, the output is often probabilistic and ambiguous, making it difficult to draw clear conclusions. These issues collectively make deepfake detectors unreliable for use as standalone tools in digital evidence workflows.

The Leinster Principle[3]

This concept describes how synthetic media achieve deception not only through technical realism, but also by intentionally layering contextual imperfections such as filters, noise, or framing that mimic the natural flaws of genuine recordings. These elements are designed to fool both human perception and AI detection systems by creating a convincing illusion of authenticity, effectively concealing the absence of genuine substance beneath.

1.2 Shallowfakes

Not all threats stem from highly sophisticated manipulations. Equally dangerous are shallowfakes, which are audio, video, or images that are crudely edited, deliberately mislabelled, or selectively framed to mislead. Unlike deepfakes, shallowfakes do **not** rely on artificial intelligence. Instead, they exploit context, timing, and selective presentation to distort meaning. Despite requiring far less technical expertise, their impact can be just as damaging, particularly when deployed in politically sensitive or emotionally charged environments. The relative simplicity of shallowfakes has led to a surge in misinformation campaigns that rely on *contextual deceit* rather than advanced computational techniques. This makes them especially difficult to regulate or detect, as traditional forensic methods may find no detectable tampering at the pixel, frame, or audio sample level.

A further evolution of this threat, examined in detail in this guideline, is the rise of hybrid content, which occupies the grey area between shallowfakes and deepfakes. These manipulations combine elements of genuine media with selectively inserted AI-generated segments, creating a new class of threat that is both technically deceptive and contextually misleading. It is estimated that in the first quarter of 2025 alone, USD 200 million was lost as a result of deepfake fraud[4]. This figure is only set to rise with the prevalence and sophistication of synthetic media improving by the day. The threat landscape extends beyond digital mischief.

Both deepfakes and shallowfakes have now been weaponized across three key domains:

1. **Personnel** where individuals are impersonated to extract sensitive information, manipulate opinions, or compromise security. The Lawrence Wong Scam case[5] demonstrates how deepfakes and shallowfakes can be weaponized to target people and personnel by impersonating a trusted public figure to manipulate opinions, and in this case, extract financial gain through deceptive, fake content.

[3] Williams, E., Rodgers, J., Jones, K., Chandler-Crnigoj, S. (2026). *Flawed by Design: The Leinster Principle and the Forensic Challenge of Synthetic Media,* [online]. Available at: https://engrxiv.org/preprint/view/6604/version/8588 last accessed 10 March 2026.
[4] Resemble.AI (2025). Q1 2025 *Deepfake Incident Report: Mapping Deepfake Incidents*. [online] Available at: https://www.resemble.ai/wp-content/uploads/2025/04/ResembleAI-Q1-Deepfake-Threats.pdf, last accessed 24 June 2025.
[5] Channel NewsAsia (2025). *PM Wong warns of deepfakes of him promoting scam products, services*. Channel NewsAsia, [online]. Available at: https://www.channelnewsasia.com/singapore/deepfake-video-prime-minister-lawrence-wong-scams-cryptocurrency-schemes-pr-services-4985111 last accessed 12 September 2025.

2. **Judicial** where synthetic media are introduced, either maliciously or mistakenly, as evidence in legal proceedings. The Pennsylvania case,[6] where the mother of a cheerleader was accused of creating deepfakes of her daughter's cheerleading rivals, illustrates how synthetic media can enter the legal sector. Whether deepfakes and shallowfakes enter proceedings due to malicious intent or mistaken assumptions, it is clear that the presence of synthetic media undermines evidentiary standards, legal proceedings, and can even go as far as to damage reputations of those involved.
3. **Operational/Security** where deepfakes or shallowfakes disrupt or exploit vulnerabilities in institutional systems. From phishing attacks using cloned executive voices to synthetic blackmail and AI-generated misinformation campaigns, these attacks often bypass technical controls and exploit human trust.

These examples demonstrate that the deepfake challenge is not simply technological; it is epistemological. How do we know what we know? How do we trust what we see or hear? And how can our systems of verification keep pace with technologies designed to evade detection? The forensic community is already grappling with this shift. Traditional methods of detection, such as frame analysis, spectral audio fingerprinting, and metadata inspection are no longer adequate when faced with AI systems that can fabricate or mimic these very patterns. Furthermore, many of the AI-powered detection tools being developed today operate as black boxes, producing outputs that cannot be easily explained, verified, or reproduced. This raises serious concerns about their admissibility in court and their reliability in forensic science, where standards of evidence must be rigorous, repeatable, and transparent. Despite these challenges, there are grounds for optimism.

Efforts are underway across law enforcement, industry, and academia to develop new standards, tools, and protocols for media authentication. Initiatives such as the Content Authenticity Initiative[7] and metadata-blockchain integration offer potential pathways to strengthening provenance and resilience against synthetic manipulation.

At the same time, training and education will be vital: for investigators, for forensic analysts, for legal professionals, and for the public. Ultimately, the deepfake era forces a reconsideration of what counts as "real". It demands both technical precision and procedural clarity. Throughout this document, we aim to provide practical tools and conceptual frameworks for addressing this complex challenge, starting with the basics of provenance, and building toward an interdisciplinary, international response that recognizes both the power and peril of synthetic media.

[6] Delfino, R. (2023). Article 3 2-2023 Proceedings from Technological Fakery, 74 HASTINGS L. *Hastings Law Journal Hastings Law Journal*, [online] 74. Available at: https://repository.uclawsf.edu/cgi/viewcontent.cgi?article=4012&context=hastings_law_journal. Last accessed 12 September 2025.
[7] Dotan, J., Hanna, S., & Gregory, S. (2020). *The Content Authenticity Initiative Setting the Standard for Digital Content Attribution 2*.

2. THREAT LANDSCAPE

Before 2022, deepfakes were more technically demanding to create than they are now. The creation process has become increasingly open source with a gradually lower barrier to entry. This fake media is now being produced with consumer-grade apps and online sites with a consistently lower cost to the creator. The UK government estimates that the number of deepfakes shared on online platforms has increased from 500,000 in 2023 to 8 million in 2025, showing an upsurge of 1,500 per cent.[8] This only serves to create an extremely attractive opportunity for criminals to take advantage – this is no longer a simple threat.

Voice-based threats have become increasingly prolific, with cases rising 650 per cent year-on-year,[9] with worldwide cases predicted to contribute up to USD 44.5 billion in global fraud annually. Law enforcement faces both volume and credibility challenges in tackling these crimes.

Additionally, the targets of these attacks have changed. Back in 2022, the risks were seen as disproportionately affecting public figures such as politicians, celebrities, and government officials, especially around the time of the US and UK elections in 2024. Now, a large proportion of deepfake victims so far in 2025 are private citizens,[10] often targeted with extortion, romance scams, or revenge pornography. This widens the policing mandate from national security to safeguarding everyday individuals – but the original threat has not disappeared.

The detection gap has widened. Generative models have become more sophisticated, with law enforcement acknowledging this issue and empirical studies showing that human observers perform only marginally better than chance when identifying deepfakes, often achieving accuracy rates of just 50-60 per cent,[11] roughly 55 per cent across all media types, with training offering limited improvement. Combining this with the liar's dividend, where genuine evidence can be dismissed as fake, this is now undermining judicial integrity.

Addressing this requires mechanisms that ensure media authenticity is considered from the outset of an investigation, rather than retrospectively, so that confidence in evidence is preserved. While some agencies may note they have not yet encountered synthetic media in their casework, this absence cannot be taken as evidence of absence. Without systematic methods to identify it, there is a risk of overlooking manipulated content. Building capacity to detect and document deepfakes at the earliest stage of the evidential chain is therefore essential.

Not only has the technology developed to create deepfake media, but the methods in which they are being deployed have evolved. With the advent of real-time synthesis, like real-time face and/or voice swapping technologies being integrated into calls, livestreams, and conferencing platforms, it is no longer just a matter of the synthetic media being the issue, but the context in which they are being used and presented.

---

[8] Accelerated Capability Environment (A.C.E) (2025) *Innovating to detect deepfakes and protect the public*. https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public. Last accessed 16 August 2025.
[9] Pindrop (2025) *2025 Voice Intelligence and Security Report - PiNDROP*. https://www.pindrop.com/research/report/voice-intelligence-security-report. Last accessed 15 June 2025.
[10] Resemble.AI (2025) *Q1 2025 Deepfake Incident Report: Mapping Deepfake incidents*. report. Available at: https://www.resemble.ai/wp-content/uploads/2025/04/ResembleAI-Q1-Deepfake-Threats.pdf.
[11] Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M. and Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538.

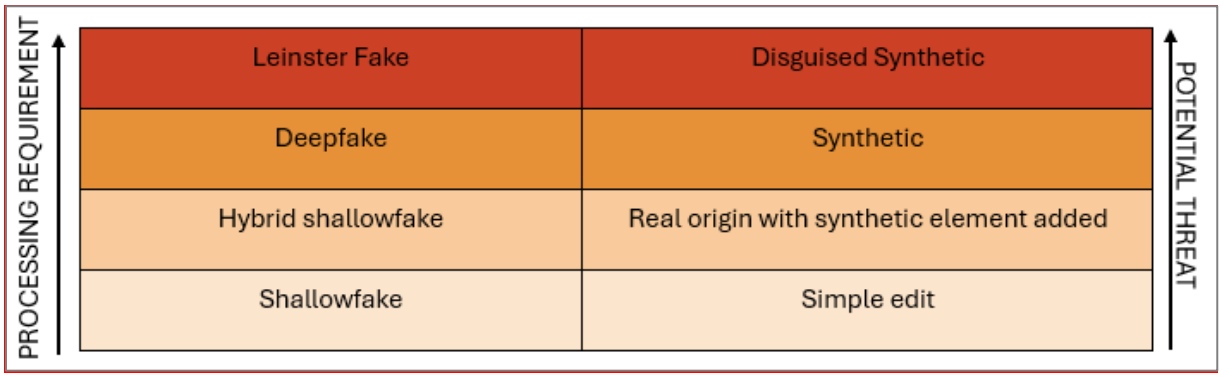

Figure 1: The deepfake severity scale

2.1 Leinster fake

In relation to the Leinster Principle, this is a deepfake, disguised to appear more realistic, lowering the threshold of believability. While deepfakes attempt to appear real, they are not always successful, but through editing after post-generation, such as applying a telephone filter, compressing the file, changing the file type, and other simple edits more synonymous with shallowfakes, they appear more believable within the context they are placed.

This is the most dangerous type of synthetic media. Not only does this mean that it is possible that the deepfake indicators can be removed, like watermarks or other detection methods mentioned in this document, but it often bypasses suspicion by blending into the context in which it was intended. This layered cake that is the Leinster Principle is the barrier that will become the most challenging to overcome in the next few years. While deepfake detection is progressing and becoming more reliable day by day, the realistic nature of the Leinster fake has yet to be addressed, since not only must the deepfake be detected, but the editing must also be identified. For current detection to work on this type of deepfake, the editing must effectively be stripped, so the detection algorithm can detect it – but how is it identified in the first place?

2.2 Shallowfakes

Definition of shallowfakes

A shallowfake is a form of media manipulation that does not utilize advanced artificial intelligence or machine learning technologies. Instead, it relies on basic editing tools to modify or fabricate content.[12] These manipulations are often described as “cheapfakes”. Within audio, this may involve changing speed or pitch, cutting and splicing speech to fabricate dialogue, or reordering words to distort meaning.

[12] Now, D.F. (2023) *Deepfakes vs Shallowfakes: Unraveling Digital Manipulation*. https://deepfakenow.com/what-is-the-difference-between-a-deepfake-and-shallowfake/ last accessed 25 August 2025.

In video, common techniques include slowing or speeding up footage, cropping scenes, rearranging sequences, or selectively omitting key visual elements to alter the context of an event. Within an image, shallowfakes can involve basic tools to remove or add objects, adjust lighting and shadows, or crop out identifying features to mislead the viewer.

All three forms depend on simple, accessible editing software and minimal technical expertise, but their impact can be profound. These edits often go undetected by casual observers and exploit a lack of contextual knowledge or confirmation bias in the audience. Crucially, unlike deepfakes, which are fully synthetic and AI-generated, shallowfakes manipulate authentic media through human intervention. The underlying content remains real, but its meaning is intentionally distorted. This makes shallowfakes especially dangerous, as they preserve an appearance of authenticity while delivering a fundamentally altered message.

2.3 Hybrid shallowfakes

Definition of shallowfakes with hybrid content

Shallowfakes with hybrid content is a relatively new phenomenon, which refers to digital media such as audio, video, or images that have undergone manual or low-skill edits, but crucially also contain limited synthetic elements generated using artificial intelligence. Unlike full deepfakes, which are entirely or predominantly AI-generated, hybrid shallowfakes combine real, authentic material with AI-enhanced or AI-inserted fragments, creating media that appear largely genuine but are partially manipulated to mislead, deceive, or distort context.

2.3.1 Audio hybrid manipulations

Within the current context, Y. Gao describes this as "segmental-level audio deepfake", which is defined as an utterance partially fabricated using both authentic and synthetic speech. Based on multiple studies and datasets,[13] detecting such hybrid audio remains challenging and under active development[14]. Similar terms in the literature include "partially spoofed",[15] while the synthetic insertion is often referred to as "voice conversion", the process of transforming one speaker's voice to sound like another without altering linguistic content.[16] This form of manipulation is part of the broader field of speech synthesis, which modifies voice identity, emotion, or accent.[17]

In audio, shallowfakes involve simple splicing, reordering, or modification of existing recordings to construct a misleading narrative. Common examples include changing intonation, inserting silences, or

[13] L. Zhang, X. Wang, E. Cooper and J. Yamagishi, "Multi-task learning in utterance-level and segmental-level spoof detection," *In Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge,* pp. 9–15, September 2021. [Online]. Available at: https://www.isca-archive.org/asvspoof_2021/zhang21_asvspoof.pdf. Last accessed 24 February 2025.
[14] Y. Gao, "Audio deepfake detection based on differences in human and machine generated speech," January 2023. [Online]. Available at: https://doi.org/10.1184/R1/21842454.v1. Last accessed 2 October 2025.
[15] L. Zhang, X. Wang, E. Cooper, J. Yamagishi, J. Patino and N. Evans, "An initial investigation for detecting partially spoofed audio," June 2021. [Online]. Available at: https://arxiv.org/abs/2104.02518. Last accessed 24 February 2025.
[16] D. G. Childers, K. Wu, D. M. Hicks and B. Yegnanarayana, "Voice conversion," *Speech Communication*, 4(5), pp. 387-396, June 1989. [Online]. Available at: https://www.sciencedirect.com/science/article/abs/pii/0167639389900411. Last accessed 11 September 2025.
[17] S. H. Mohammadi and A. Kain, 2015, "An overview of voice conversion systems," *Speech Communication*, 27(2), pp. 1-14. April 2017. [Online]. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0167639315300698. Last accessed 10 September 2025.

trimming words to suggest someone said something they never did. These manipulations are often referred to as “cheapfakes” and require no artificial intelligence, only basic editing tools.

Hybrid audio shallowfakes introduce synthetic elements into genuine recordings. A common technique is voice conversion, where a small segment like a word or sentence is generated using a cloned voice and spliced into the original clip. These segmental-level audio deepfakes are particularly difficult to detect because they blend real and synthetic speech seamlessly. Even trained listeners may not notice the alteration, and many detection systems may pass such content as genuine unless subjected to in-depth forensic analysis. The absence of the original unedited audio complicates the verification process, shifting the burden onto the examiner to prove what was artificially inserted or removed.

2.3.2 Video hybrid manipulations

In videos, hybrid shallowfakes may involve authentic footage that has been selectively altered using AI-based stabilization, lip-synchronization tools, or facial-blending filters. These enhancements often remain below the threshold of full synthetic generation, making them difficult to detect using deepfake-focused tools. When combined with conventional edits such as cropping, reordering, or context removal, these subtle modifications can misleadingly alter meaning while retaining visual realism.

Visual shallowfakes often rely on simple editing techniques like cropping, trimming, or reordering footage to distort context. In protest or conflict scenarios, videos can be subtly altered to imply provocation, aggression, or to conceal misconduct. These edits are deceptively effective because they manipulate authentic footage rather than fabricating new content.

Hybrid video shallowfakes take this further by incorporating AI-based enhancements without becoming fully synthetic. For example, a shaky handheld recording might be stabilized using machine learning tools, or facial expressions could be subtly modified with commercial-grade filters to appear more convincing. While these additions do not qualify as deepfakes, they significantly affect how content is perceived and can hinder forensic review. Because many detection tools are calibrated to flag fully synthetic media, they may overlook these hybrid forms, particularly when there is no original footage available for comparison.

2.3.3 Image hybrid manipulations

For images, shallowfake hybrids may blend genuine photographs with minor AI-generated alterations, such as facial expression changes, background replacements, or light retouching through tools like generative fill. The result is an image that appears largely authentic but has been deliberately modified to deceive or mislead. As with audio and video, these hybrid manipulations complicate forensic analysis and challenge existing authentication protocols.

Shallowfake images rely on conventional editing tools to subtly manipulate visual context. This could involve cropping out crucial elements, changing contrast or saturation to obscure features, or combining elements from multiple photos into a single misleading composite. These low-tech changes are especially effective in surveillance, identification, or evidentiary settings, where even minor edits can influence perception or introduce doubt.

Hybrid shallowfake images may incorporate generative techniques like AI-powered facial enhancement, background replacement or blemish removal. These are often deployed to hide signs of manual

tampering, making the image appear more natural to viewers and harder to flag in forensic workflows. Despite not being entirely AI-generated, these modifications can obstruct the authentication process, especially when metadata have been stripped or altered.

Other descriptors, such as “partial fakes” or “composite fakes”, have also been proposed to describe these phenomena.

What’s the difference: Context vs content

- **Shallowfakes** primarily alter the context by manually editing how real content is presented, such as through cutting, splicing, rearranging, or changing playback speed. No AI is used.
- **“Hybrid shallowfakes”** can be thought of as manipulation to both context and content, combining manual edits with AI-generated, spliced-in elements.

Specific examples pertaining to each media type can be seen above.

These forms of manipulation can be difficult to distinguish, especially when the media appear largely authentic. However, it is important for law enforcement professionals to be aware of the minute “differences” and to recognize that even minimal or subtle alterations can influence how content is interpreted. While it might be tempting to dismiss such material as entirely fake or manipulated, it is important to recognize that these recordings, videos, or images often contain significant genuine content alongside the synthetic inserts. This authentic material can be critical as evidence, even when parts have been artificially introduced. Developing an understanding of these techniques and anticipating the possibility of nuanced or low-level manipulations support more effective assessment and handling of digital media evidence.

2.4 Impact on law enforcement

These layered manipulations, whether in audio, video, or images, create a forensic grey zone. Hybrid shallowfakes often bypass traditional detection methods, leaving practitioners to prove what has been changed rather than identify what is outright false. This challenge is compounded in legal and institutional contexts where source material is unavailable, and authenticity must be judged solely from the suspect file.

In addition, “stream recording” or “media replay”, which is the practice of recording media during playback rather than accessing or preserving the original file, further undermines forensic analysis by stripping metadata, reducing quality, and flattening unique digital signatures. This process often leaves only a degraded copy, obscuring signs of manipulation and complicating efforts to verify authenticity.

Raising awareness of these evolving threats is a critical first step, and understanding how practices like stream recording compromise verification processes can help institutions, practitioners, and the public better navigate today’s complex media environment. Addressing these complexities will require sustained collaboration with researchers and the academic community to develop adaptive detection tools, robust standards, and shared frameworks for verifying truth in an increasingly synthetic media landscape.

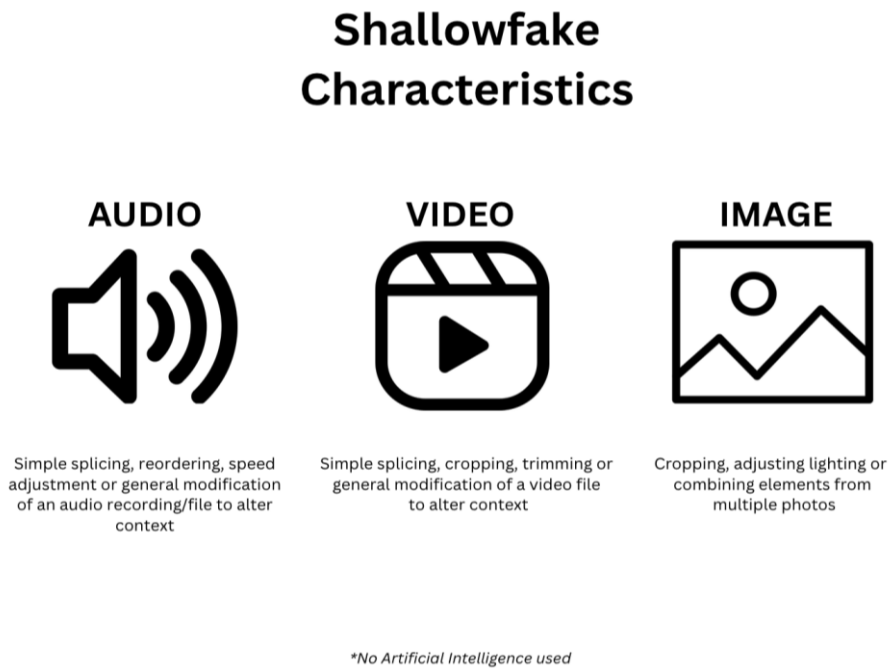

Figure 2: Characteristics of a shallowfake as presented through audio, video, and image

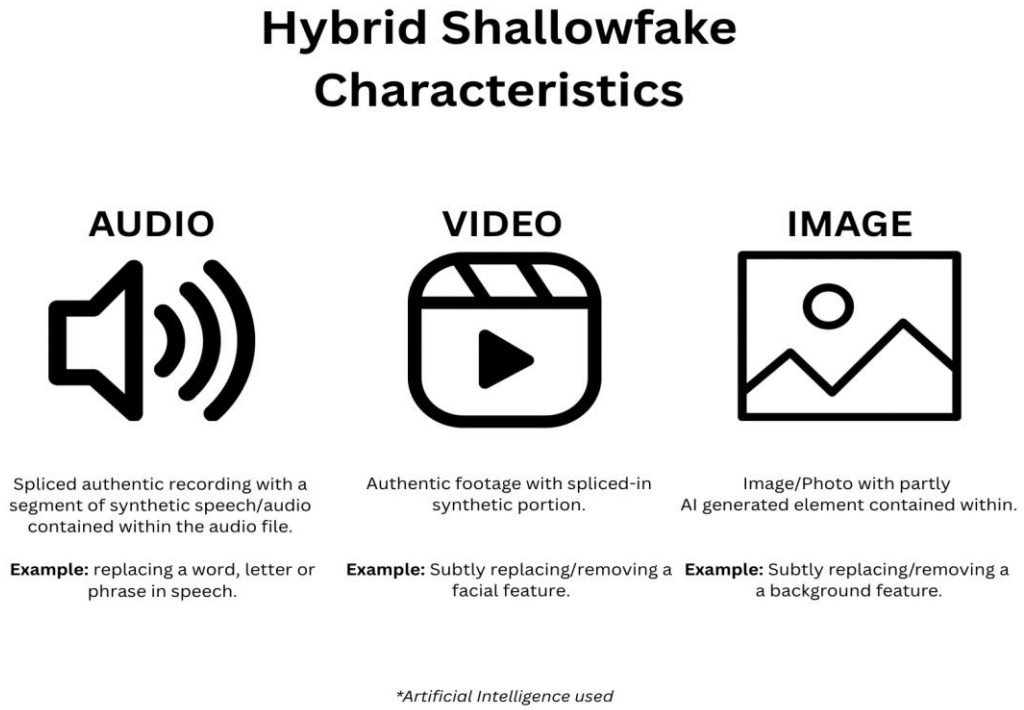

Figure 3: Characteristics of a hybrid shallowfake as presented through audio, video, and image

2.5 Real-world cases

Several notable cases in recent years highlight the growing impact of both shallowfake and AI-generated media on public perception, reputations, and legal proceedings.

In 2019, a widely circulated video of US House Speaker Nancy Pelosi was deliberately slowed down and pitch-altered to create the false impression that she was intoxicated.[18] Despite the crude nature of the edit, the video gained significant traction online before social media platforms intervened. This case remains one of the most well-known examples of shallowfake video and audio manipulation, demonstrating how basic edits can have disproportionate influence.

In 2021, the FBI issued a warning that synthetic media would likely be exploited for cybercrime and foreign influence in the near future, a prediction that was realized in 2025. That year, Eric Eiswert, former principal of Pikesville High School in Baltimore, USA, filed a lawsuit after being suspended in connection with a racist audio recording attributed to him. Subsequent forensic analysis, including input from an FBI-contracted specialist, determined that the recording had been artificially generated using AI, with human edits added post-creation to insert background noise and simulate authenticity[19].

The case illustrates how AI-generated media, even when blended with real elements, can lead to serious reputational and legal consequences.

Collectively, these examples demonstrate the increasing complexity faced by forensic practitioners. Whether dealing with entirely synthetic media, shallowfakes, or hybrid manipulations that blend authentic and artificial elements, investigators must navigate a rapidly evolving landscape. The growing sophistication of these techniques challenges traditional authentication methods and highlights the urgent need for continued development of forensic tools, training, and legal frameworks to ensure reliability and trust in audio evidence, particularly within criminal justice and institutional contexts.

The future

Shallowfakes are fast and scalable; unlike deepfakes, which require extensive datasets and computing power, shallowfakes can be created and spread rapidly across social media and messaging platforms. Their reach can be amplified by misinformation actors, political campaigns, and bots, making them effective disinformation tools. Legally, shallowfakes create uncertainty in courts and evidence chains. Edited media may appear authentic at a glance, yet small changes can significantly alter meaning. The growing suspicion around manipulated media also fuels the "liar's dividend", where genuine evidence is denied based on claims of fakery, undermining trust in digital content broadly.

In this environment, the concept of a "truth tax" becomes ever present: a growing cost, whether financial, technical, or cognitive imposed on those seeking to verify what's real. As misinformation spreads faster and verification becomes harder, the burden of proving authenticity increasingly falls on law enforcement agencies and journalists.

[18] BBC News (2019). *Nancy Pelosi clip shows misinformation still has a home on Facebook*. BBC News, [online]. Available at: https://www.bbc.co.uk/news/technology-48405661. Last accessed 12 August 2025.
[19] AI Incident Database (2024). *High school athletic director in Baltimore County allegedly created racist deepfake audio impersonating principal.* AI Incident Database, [online]. Available at: https://incidentdatabase.ai/reports/3837/. Last accessed 1 September 2025.

Regulatory responses to manipulated media have primarily centred around AI-generated deepfakes, despite the fact that comprehensive legal frameworks are still in development in most jurisdictions. Existing proposals and emerging regulations, such as the EU AI Act[20] or draft laws in the US[21] and UK[22] have largely focused on addressing the risks associated with generative AI, with lower-tech forms of manipulated media, such as shallowfakes, receiving comparatively less attention

Generally, around the world, there is a marked lack of direct legislation around the use of AI and the creation of harmful synthetic media. However, there appears to be indirect legislation in place that may be used in the case of synthetic media – fraud, defamation, copyright infringement, and sexual content have implications on the creation of these files, though synthetic media are yet to be widely recognized as an issue within their own right.[23]

This creates a critical gap, as shallowfakes can be equally damaging, particularly in sensitive domains such as criminal justice, national security, and governmental discourse. To address this oversight, legal and policy measures must be broadened to include deliberate contextual alterations achieved through basic editing tools, which are often easier to produce and harder to detect than fully synthetic content.

Addressing the threat of shallowfakes and hybrid manipulations requires a multipronged strategy: training legal and investigative professionals to recognize both conventional and AI-assisted edits; advancing detection methods that go beyond synthetic content to include manual tampering; and implementing robust provenance protocols to trace the origin, editing history, and custody of media files. Such measures are essential not only for ensuring the reliability of digital evidence but also for maintaining public trust in an increasingly manipulated information environment.

In summary, shallowfakes pose a significant yet underappreciated threat. Their simplicity, realism, and resistance to detection make them powerful misinformation tools, especially in legal, political, and institutional settings. Attention must be paid to these low-tech manipulations as synthetic media detection evolves.

[20] European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union, L 2024/1689, pp. 1-144.
[21] S.2938 - 119th Congress (2025-2026): Artificial Intelligence Risk Evaluation Act of 2025. (2025, September 29). Available at: https://www.congress.gov/bill/119th-congress/senate-bill/2938.
[22] Artificial Intelligence (Regulation) Bill [HL], (2025). Available at: https://bills.parliament.uk/publications/59353/documents/6094#:~:text=(1)%20This%20Act%20extends%20to,Intelligence%20(Regulation)%20Act%202025.
[23] Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, *100*(3), 1275–1286. Available at: https://doi.org/10.1093/ia/iiae073.

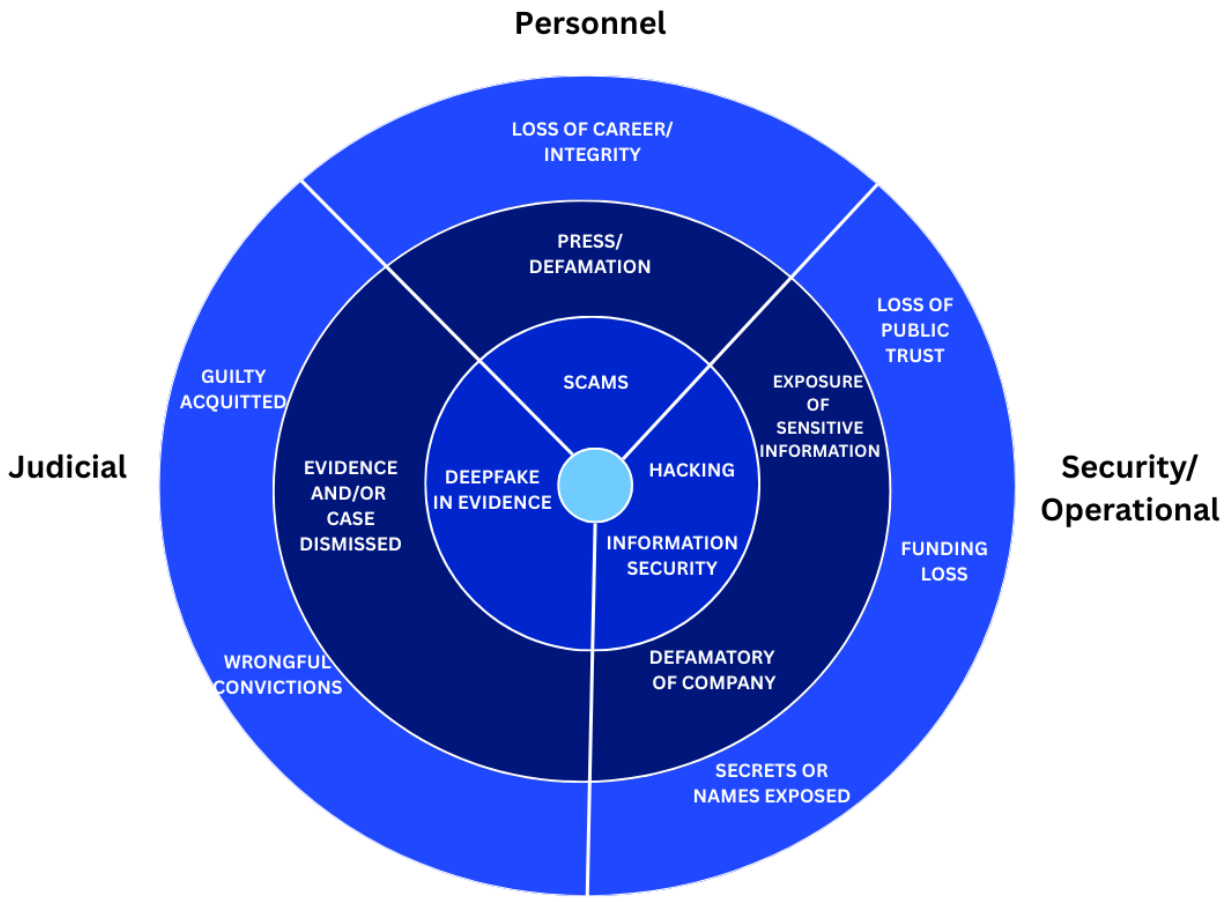

Figure 4: Three main areas of threat with the immediate risks in the centre, radiating outwards in time scale

To effectively demonstrate the depth and diversity of the deepfake threat, three case studies will be presented, each highlighting a different sector impacted by this technology.

The Lawrence Wong scam case

The Lawrence Wong scam case underscores the far-reaching and deeply personal consequences of deepfake technology. In this case, AI-generated videos and voice clones falsely portrayed Singapore Prime Minister (PM) Lawrence Wong promoting fraudulent schemes[24]. This "fabricated authority" meant scammers could manipulate the public. While PM Wong responded to the circulating deepfake through a verified Facebook post, official communication channels may face challenges in effectively countering manipulated content since messages are overlooked, misinterpreted, or viewed with suspicion.

Importantly, the harm caused by deepfakes in this way extends beyond their direct targets. Public figures, like PM Wong or US Senator Marco Rubio, in a similar reference case, may be the face of the attack, but the victims also include members of the public who are misled, exploited, or financially harmed, making

[24] Channel NewsAsia (2025). *PM Wong warns of deepfakes of him promoting scam products, services*. Channel NewsAsia, [online]. Available at: https://www.channelnewsasia.com/singapore/deepfake-video-prime-minister-lawrence-wong-scams-cryptocurrency-schemes-pr-services-4985111 last accessed 12 September 2025.

them secondary victims of misinformation.[25] Deepfakes can slander careers and damage reputations, ultimately destroying livelihoods. These cases are not limited to national leaders; even more localized figures, such as the Salt Lake City Police Chief, have been targeted in sophisticated scams where perpetrators, once again, achieved tangible results.[26] These incidents reveal a disturbing truth, which is that anyone in a position of power or public trust is vulnerable, and the cumulative effect is an erosion of societal confidence in authority figures and members of public institutions.

The Pennsylvania Cheerleader Mom

The "Pennsylvania Cheerleader Mom" case shows the growing danger that the prevalence of deepfakes can pose to the integrity of the legal system, not just through their creation but through the very possibility of their existence. The 2021 case features a Pennsylvania mother, Raffaela Spone, who was charged with harassment after being accused of creating deepfake images and videos to frame her daughter's cheerleading rivals. These visuals allegedly showed the teens drinking, vaping, and appearing nude. The case made national headlines in the United States; however, prosecutors later withdrew the deepfake-related claims when it was revealed that law enforcement based their conclusions on a mere visual inspection and what the lead detective called a "naked eye" assessment[6].

No forensic analysis was conducted, and no expert validation supported the claim that the media were synthetic. This highlights the critical situation in the legal system with the vulnerability of judicial proceedings regarding the prevalence of deepfakes. Not only can deepfakes be used to fabricate evidence, but the *threat* of deepfakes can also trigger flawed legal actions such as false accusations. In this case, subjective interpretations by police and prosecutors were substituted for rigorous evidentiary and forensic standards. This raises alarms about how the justice system will cope with increasingly convincing synthetic media because, as deepfakes become harder to detect and easier to create, the courts and police face a dual threat. The threat is being deceived by fakes or misidentifying real content as fake. Both outcomes erode public trust and reveal the urgent need for legal reform and set-in-stone standards of forensic and non-forensic authentication processes.

The Hong Kong bank incident

The Hong Kong bank incident demonstrates the alarming sophistication with which deepfakes can be used to exploit the security of institutions. In this case, a finance worker at the multinational firm in Hong Kong, China was deceived into authorizing the transfer of approximately USD 25.6 million after attending a video call where every participant, including a convincing replica of the company's UK-based CFO, was a deepfake.[27] Initially sceptical after receiving a suspicious message about a confidential transaction, the employee was ultimately convinced by the seemingly authentic appearance and voices of colleagues during the call. Unbeknownst to the employee, all participants were AI-generated fakes. According to Hong Kong police, this scam was part of a broader pattern involving stolen identities, fraudulent loan applications, and more than 50 fake bank accounts, all of which leveraged deepfake technology to bypass

[25] The Guardian (2025). *AI scammer posing as Marco Rubio targets officials in growing threat*. The Guardian, [online]. Available at: https://www.theguardian.com/us-news/2025/jul/08/marco-rubio-ai-impostor. Last accessed 1 December 2025

[26] KSL (2024). *Deepfake of Salt Lake police chief prompts warning about AI scams impersonating police*. KSL, [online]. Available at: https://www.ksl.com/article/51163390/deepfake-of-salt-lake-police-chief-prompts-warning-about-ai-scams-impersonating-police?

[27] Lin, L. S. F. (2025). Examining the role of deepfake technology in organized fraud: Legal, security, and governance challenges. *Frontiers in Law*, *4*, 6-17. https://doi.org/10.6000/2817-2302.2025.04.02. Last accessed 11 November 2025.

facial recognition systems. The case was only uncovered after the employee sought verification from the company's headquarters, by which time the funds had already been transferred.

This incident reveals how AI-generated deception can convincingly simulate human presence, bypass institutional safeguards, and compromise the very foundation of trust that underpins modern finance and legal compliance. The implications are profound: voice and video, once pillars of personal verification, can no longer be taken at face value. As law enforcement agencies struggle to adapt, and detection tools lag behind creation tools, financial institutions and legal systems face an urgent need to reevaluate authentication protocols, liability frameworks, and regulatory protections in the age of AI fraud.

These cases only highlight the impact that synthetic media can have on many aspects of daily life, not just in policing. Though it seems we are now seeing a new threat emerge to staff themselves. Simply attending an online call could result in a voice sample being taken of anyone who speaks, creating a synthetic clone, and impersonating them, thus putting their professional credibility at risk should this be used maliciously. The risks have now become enormous, posing a serious threat not only to the justice system, but the security of policing institutions and government bodies themselves.

3. PREVENTION WITHIN CHAIN OF CUSTODY

In the fight against synthetic media and manipulated digital content, provenance is one of the most critical yet often underemphasized pillars of prevention. Provenance, in this context, refers to the origin, ownership history, and handling record of a piece of media from creation to submission as evidence or intelligence. Just as biological evidence is handled under a strict chain of custody to avoid contamination, so too must digital media be managed to maintain their integrity and trustworthiness.

Establishing and maintaining a verifiable chain of custody is not merely a bureaucratic necessity. In a digital environment where deepfakes and shallowfakes can be produced and disseminated with ease, provenance often provides the first and sometimes only layer of defence against manipulation. This section explores how provenance can be preserved through technical, procedural, and forensic means, and how the absence of provenance can undermine entire investigations or judicial processes.

3.1 The importance of provenance within a synthetic media landscape

The rapid proliferation of deepfake technology has shifted media forensics in general into a high-stakes arena. Where once we could rely on the assumption that a piece of audio, photograph, or video represented a truthful moment in time, today we must interrogate every file's authenticity before it is accepted as evidence.

Deepfakes erode public trust, not only in the content they manipulate but in the systems meant to distinguish real from fake. In this context, establishing provenance early and consistently is critical. Provenance provides:

- **A contextual framework**: When were the media created? By whom? With what device?
- **Tamper detection points**: Has the file been modified since its creation?
- **Legal admissibility**: Is the file's chain of custody documented and defensible?
- **Multi-source corroboration**: Cross-referencing with other data to strengthen evidentiary consistency.

If provenance cannot be verified, even authentic media may be rendered unusable due to doubt and lack of evidentiary weight.

In a world where anyone can claim that a video is not real, it is provenance and not just forensic analysis that can potentially settle the matter. In contrast to conventional digital tampering, deepfakes and shallowfakes present unique challenges to provenance because their synthetic origins often simulate legitimate creation processes.

While classic digital manipulation typically leaves behind detectable editing traces, AI-generated or subtly manipulated content may be designed to look "native", with no obvious alteration indicators. In such cases, the absence of reliable provenance becomes not just an investigative inconvenience but a potential vector for disinformation, false accusation, or legal derailment.

For global policing and justice systems, this means provenance is not only a desirable attribute, but increasingly a non-negotiable baseline**.** Without it, we may have no way to distinguish truth from imitation.

3.2 Evidence of manipulation – Provenance as detection

During forensic inquiry of a file, due forensic process requires the provenance of the evidence to be ascertained as far as possible. Before specific techniques are undertaken for deepfake or synthetic media detection, identifying possible areas of manipulation in the metadata or the evidence contents themself can bring doubt upon an article of evidence, and mark it as a possible synthetic file.

3.3 Technical approaches to provenance

3.3.1 Metadata preservation and validation

Metadata include timestamps, device information, GPS coordinates, codec specifications, and more. When preserved accurately, this embedded information can offer crucial insight into:

- whether the file's claimed source and contents align.
- whether it has been edited, transcoded, or manipulated.
- which tools were used in its creation or modification.

However, metadata can be stripped or forged. For this reason, investigators must rely on secure file transfer methods and immediately document or extract metadata before any file conversion or processing occurs.

Tools such as ffmpeg, ExifTool, and proprietary forensic suites can assist in extracting and validating metadata. But it is equally important to validate the consistency between metadata and file behaviour, for example, checking if an audio file's bitrate is consistent with the metadata's stated device.

3.3.2 Hashing and digital watermarking

"Hashing" similarly provides a file "fingerprint" at any point in the chain of custody. If the hash changes, it indicates the file has been altered in some way. Watermarks and hashes are particularly effective when implemented at the point of capture (e.g. on police bodycams or surveillance systems).

Digital watermarking involves embedding invisible markers into files at the point of creation. These markers can confirm a file's source and integrity if later challenged.

In practice, many proprietary tools will watermark the files they produce to prove it originated from their software or to identify unauthorized copies. This has parallels with copyright infringement protections, where watermarking helps identify fraudulent or misattributed material.

However, the use of watermarking has limitations. If the method by which it is conducted becomes widely known, there is a possibility that the mark could be inserted into content that did not originally contain it

or removed from content that did. In the context of synthetic media within the legal system, there will undoubtedly be the question of proving or disproving that a watermark has been added or removed on any suspect material, and what degree of certainty should be required before an investigator treats a watermark as trustworthy evidence.

Considering this, digital watermarking of synthetic media at the point of creation should not be relied upon as a standalone solution, but as a component of a broader provenance framework.

3.3.3 Blockchain and distributed ledger technologies (DLT)

Blockchain is a type of database that records information in a way that makes it extremely difficult to tamper with or change. Instead of being stored in one place, records (or "blocks") are shared across a network of computers and linked together in a secure, chronological chain. This makes it useful for tracking the history, or provenance, of digital files.

One area where blockchain is being explored is content verification, particularly in journalism and legal evidence handling. By logging every interaction with a file in a secure, tamper-resistant ledger, you can create an unchangeable "chain of custody" that shows exactly where the file has been and who has handled it. However, blockchain is not a magic fix. It only works reliably when it is built into the process from the moment a file is created, not added later. And it should be used alongside traditional forensic methods, not as a replacement.

In parallel, the Content Authenticity Initiative (CAI) [28], led by Adobe and supported by a range of media and technology companies, is working on open standards to help verify the origin and history of digital content. It uses tools like digital signatures and cryptographic timestamps to attach trustworthy metadata to photos, audio, videos, and documents. These metadata travel with the file, enabling end-users to trace its origins and assess whether it has been altered.

The CAI is also a founding member of the Coalition for Content Provenance and Authenticity (C2PA), which develops the technical specifications that power these authenticity tools. The C2PA standard defines how metadata about a file's creation, editing, and distribution are securely embedded and verified.[29]

In essence, this can be thought of as a new "provenance technology" that provides a new framework for establishing content provenance across platforms and formats. This technology can help "explain whether a photo was taken with a camera, edited by software, or produced by generative AI."[30]

3.4 Procedural best practices

[28] Content Authenticity Initiative. (2025). *Content authenticity initiative*. Content Authenticity Initiative. https://contentauthenticity.org/how-it-works. Last accessed 19 October 2025.
[29] C2PA Technical Working Group. (2025). C2PA Content Credentials Explained: Addressing Common Questions and Updates. https://c2pa.org/wp-content/uploads/sites/33/2025/10/content_credentials_wp_0925.pdf.
[30] Richardson, L. (2025) 'How we're increasing transparency for gen AI content with the C2PA,' Google, 21 July. https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/.

3.4.1 Chain of custody documentation

Each individual who handles the file should be recorded, along with the date, time, purpose of access, and tools used. Standardized digital evidence forms or integrated evidence management systems can enforce these protocols.

Institutions should enforce minimum chain of custody requirements before allowing files to enter forensic workflows. These may include:

- a verified source statement.
- a log of all handling steps.
- integrity verification (e.g. hash matching) at each stage.

Forensic science is grounded in the principle of repeatability and transparency. Provenance fits into this by allowing:

- validation of whether the file should be analysed at all.
- identification of suspicious origin claims before wasting resources.
- prioritization of analysis based on authenticity likelihood.
- documentation that can be defended in court if questions arise.

In a typical audio, video, and image forensic workflow, provenance serves three main functions:

1. Gatekeeping: Files with unverifiable or compromised provenance may be excluded.
2. Contextualizing: Understanding what the file claims to show and how that relates to its source.
3. Corroborating: Cross-referencing metadata and source logs against analytical findings.

3.4.2 Detecting evidence of manipulation through provenance

One of the most overlooked tools in deepfake detection is provenance analysis, examining the origin, history, and metadata of a digital file. Even before applying forensic techniques to the content itself, inconsistencies in metadata or file structure can flag potentially manipulated media. Common indicators can include:

- Creation or modification dates that post-date the claimed capture date.
- Mismatches between the listed camera or device model and file format.
- Microphone specifications are inconsistent with the analysed audio signal.
- Files with missing or stripped metadata, suggesting potential obfuscation.
- Duplicate hashes appearing in unrelated submissions hinting at reused or altered evidence.
- Compression artefacts or encoding signatures inconsistent with the claimed source device or platform.
- Comparison with exemplar recordings, authentic samples from the same device or environment, that help reveal anomalies through device-specific fingerprints or noise patterns.
- GPS or location tags that contradict the stated context.

While not definitive proof of manipulation, these red flags provide early signals that a file may warrant deeper forensic scrutiny. Provenance analysis, when integrated into the verification process, helps build a more robust chain of trust for digital evidence.

3.5 International jurisdictional differences in provenance handling

While both deepfakes and shallowfakes are not confined by borders, standards for provenance handling vary significantly between jurisdictions. Audiovisual material may be admissible under specific evidentiary conditions in one country, while in another it may not meet the threshold for forensic consideration at all. These differences can create challenges in cross-border investigations, especially where synthetic media are involved and visual or auditory realism can no longer be assumed to reflect authenticity.

Without coordinated standards for provenance and, more foundationally, detection and authentication, legal processes around the world become vulnerable to manipulation.

Such regional inconsistencies do not just obstruct cross-border investigations, but they also pose a significant risk to international security. Malicious actors can exploit these jurisdictional differences to spread synthetic media from regions with lesser controls, and they can target institutions in areas with strict regulations. The presence of deepfakes and shallowfakes, without coordinated, international standards, complicates joint operations between states and erodes trust in shared intelligence or digital evidence.

4. DETECTION

Human perception is fallible when it comes to deepfake and synthetic media detection. In one study, "participants perceived the identity of an AI-generated voice to be the same as its real counterpart approximately 80 per cent of the time, and correctly identified a voice as AI generated 60 per cent of the time"[31]. Another study demonstrated that regardless of having forensic expertise, one third of the time deepfake files were correctly identified, one third of the time deepfake files were identified as real and one third of the time real files were identified as fake[32].

This chapter introduces some of the current signal-based techniques and forensic analyses used to detect deepfake media. Several of these approaches, such as edge detection, pixel analysis, and compression forensics, have long existed in digital media forensics and are now being adapted to meet the challenges posed by synthetic media. This is not a sudden emergence of brand-new methods, but a retooling of established practices in response to a rapidly evolving threat. As generative AI technologies advance, so too do adversarial techniques aimed at avoiding detection[33], with active and ongoing research by malicious actors pushing the boundaries of what can be faked and hidden.

For this reason, no single technique should be relied upon in isolation. These tools must be applied as part of a suite of complementary measures, including provenance tracking, metadata analysis, contextual evaluation from law enforcement and intelligence communities, and the use of legal or regulatory frameworks where applicable. Detection should be guided by a holistic view that combines technical analysis with situational awareness, using confidence thresholds rather than binary judgments to assess the likelihood of manipulation.

Occasionally, there tends to be a strong bias towards focusing on video and image analysis, likely because visual media dominates social media platforms and much of the forensic evidence from sources like CCTV and bodycams is video-based. However, audio analysis remains comparatively underexplored and appreciated. Given that audio manipulation can be just as convincing and sometimes easier to overlook, it is crucial that more attention and resources be directed to developing robust audio forensic techniques to complement visual detection efforts. Generally speaking, current detection methods hinge on the suspect file being "abnormal" or inconsistent with the identifiers of real media.

The list of techniques is not exhaustive, but an initial insight into the possibilities. While some organizations may choose to protect their internal processes for security purposes, there is research to suggest that there are many varied, multimedia focused practices that enable detection, without the use of AI or machine learning (ML) to identify or authenticate these files.

Staying ahead requires ongoing collaboration between researchers, law enforcement agencies, and government from the outset.

---

[31] Barrington, S., Cooper, E.A. and Farid, H. (2025) 'People are poorly equipped to detect AI-powered voice clones,' Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-94170-3 . Accessed 1 December 2025.

[32] Rodgers, J., Jones, K. O., Robinson, C., Chandler-Crnigoj, S., Burrell, H., & McColl, S. (2024). Evaluating the Threshold of Authenticity in Deepfake Audio and Its Implications Within Criminal Justice. *Proceedings of the SLIIT International Conference on Engineering and Technology, 2024*, 239–248. https://rda.sliit.lk/handle/123456789/3792. Accessed 16 September 2025.

[33] Zhou, Z., Sun, K., Chen, Z., Kuang, H., Sun, X., & Ji, R. (2024). StealthDiffusion: Towards Evading Diffusion Forensic Detection through Diffusion Model. *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, 3627–3636. Available at: https://doi.org/10.1145/3664647.3681535.

4.1 Forensic approach

Detection methodologies within forensic science can be split into two main activities:

a) Identification

The initial identification of an article as potentially containing synthetic media raises a red flag, which must then be confirmed through authentication. The identification process however, not necessarily forensic, may be raised by anyone – legal defence team, prosecution team, bystander, witness, or forensic practitioners, for example. Depending on the context and allegation by which this is made, some identifiers will mean these suspect articles will need to be sent for synthetic media authentication. This step of identification is where current AI and ML methods of deepfake detection are currently more acceptable, as they do not provide a definitive yes or no answer, but do in fact provide a likelihood factor that can be followed up on in the authentication steps to follow.

b) Authentication

Authentication is the full forensic processing of suspect files, from which a more definitive forensic opinion can be formed regarding the authenticity of the file based on currently established markers and methodologies. Within this, there are markers that have been identified for certain types of synthetic media which can be found below. In most workflows, authentication is reserved for files that have been previously identified as suspect, ensuring efficient use of forensic resources.

One current and ongoing debate is that of AI and ML being included in forensic workflows. There are many such arguments for and against; however, this should be decided upon within the jurisdiction of each country. While AI and ML processes have their merits in saving time and occasionally institutional finances, there come inherent risks when framed within forensic science, such as repeatability of the processing, explainability, and the knowledge of the examiner utilizing the tool. Within the EU, it is required that those deploying AI tools are able to explain how the tool is using the data.[34] This is applicable within forensic science.

For these methods to be considered forensically viable, they must be based in science and proven to be measurable, repeatable, and peer reviewed, and therefore in keeping with forensic science methodology. Any future detection methods must also prove the same.

[34] Office of the European Union L-, P., & Luxembourg, L. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance.* Available at: http://data.europa.eu/eli/reg/2024/1689/oj.

4.2 Video/Image

a) Signal-based detection techniques

Edge detection

A computer vision technique that identifies and highlights boundaries where image properties change. In genuine videos and images, these edges are normally identified visually as the edges of objects and features such as wrinkles, hairlines, or lip/eye contours. In deepfake detection, abnormalities in these edges can be telling of its generation – blurred or softened edges where a natural feature would have a hard edge and edge discontinuities, where stitching or blending occurs between generated regions and real backgrounds, indicate synthetic media.

Specifically in video where many sequential images are being dealt with, inconsistent edge patterns across each frame that do not match with feature movements, appearing to flicker across time, indicate that synthetic media are present. This is a commonly used detection method, especially within deep-learning-based algorithms used for analysis such as convolutional neural network (CNN) filters. Classical (non-AI or ML) versions of this technique are known as:

- Sobel filter – A quick method of identifying edges in an image by looking at where the brightness/intensity of the pixels changes most sharply both vertically and horizontally.
- Canny filter – A more advanced method than the Sobel filter. It smooths the image to reduce noise, then finds where edges are most likely to be (a process similar to Sobel), then cleans the identified edges so they are thin and continuous.
- Laplacian filter – Targets places in the image where brightness changes very suddenly in all directions at once. A highly sensitive technique that can highlight very small details; however, it often results in noise being picked up, so it is often combined with smoothing to make it more stable.

Application in images: Edge detection in images involves identifying sharp changes in intensity or colour to outline objects and features. In deepfake images, unnatural edges around facial features or inconsistencies in the background can indicate manipulation.

Application in videos: In videos, edge detection is applied frame by frame to reveal irregular transitions or inconsistencies in motion boundaries. Deepfake videos may display flickering or unstable edges around faces, especially during head movements or facial expressions.

b) Pixel analysis

Images: Pixel analysis in images focuses on detecting unnatural pixel arrangements, blending issues, or irregular textures. Deepfake images often contain localized artefacts or colour mismatches at the pixel level that deviate from natural image patterns.

Videos: In videos, pixel analysis tracks pixel behaviour across frames to spot inconsistencies. Temporal artefacts, such as unnatural pixel transitions or inconsistent lighting across frames, can signal synthetic content in deepfake videos.

c) Residual noise

Images: Residual noise in images refers to the statistical noise left after removing the main image content. Authentic images exhibit consistent sensor noise, while deepfakes may lack this or show irregular patterns due to generative models.

Another cue lies in "**shadow vanishing points"**. In real images, shadows typically align toward consistent vanishing points based on the light source. In deepfakes, these may be physically inconsistent or misaligned, exposing the scene as artificially generated.

Videos: For videos, residual noise analysis involves checking frame-to-frame consistency of noise patterns. Deepfake videos often have altered or smoothed noise due to frame synthesis or compression, which can differ from naturally recorded footage. Inconsistencies in lighting and shadow behaviour over time, including unstable or non-physical shadow vanishing points, may also indicate manipulation.

d) Group of Pictures (GOP) analysis

Videos: GOP in videos determines how frames are structured, combining keyframes and predictive frames. Deepfakes may disrupt the natural GOP pattern, such as unusual keyframe spacing or altered motion prediction, revealing tampering.

e) Compression analysis

Images: Compression analysis for images examines artefacts from JPEG or other image compression formats. Deepfake images may display abnormal compression noise, inconsistent quality in different areas, or unusual quantization tables due to generation or post-processing.

Videos: Within videos, compression analysis looks at how codecs process frames and detect inconsistencies introduced by deepfake generation. Anomalies in bitrate distribution, block artefacts, or motion vector inconsistencies can indicate synthetic editing or recompression.

f) Other signal-level clues

Colour inconsistencies: Deepfakes sometimes show mismatched skin tones, unnatural gradients, or colour bleeding between the face and background due to poor blending.

Lighting inconsistencies: Lighting direction and intensity may not match the scene or background, revealing synthetic face insertions or rendering mismatches.

g) Specular reflection errors

Specular highlights, such as reflections in the eyes or on the skin, may appear misplaced or missing in deepfakes, since models often fail to simulate realistic light interactions.

h) Temporal inconsistencies

Facial feature instability: In deepfake videos, facial elements such as the eyes, mouth, or nose may shift position unnaturally between frames. These misalignments are due to the frame-by-frame generation process.

Motion inconsistencies: Natural head and body movement should follow biomechanical constraints. Deepfakes might show unnatural stiffness, jittery motion, or inconsistencies between facial expression and body posture.

Body/Face coordination errors: In some deepfakes, the face may move independently of the rest of the body, or appear detached during fast movements, revealing poor spatial integration.

i) Biological and behavioural cues

Eye blinking patterns: Deepfake models may fail to replicate natural blinking frequency or duration. Unusual blinking patterns, such as rarely blinking or too-regular blinking, can be a sign of synthetic generation.

Lip sync accuracy: Authentic videos typically show precise synchronization between lip movement and speech. Deepfakes often struggle to maintain accurate lip sync, especially with complex or emotional speech.

Gaze direction and eye movement: Real humans have varied and naturally coordinated eye movements. Deepfakes may exhibit unnatural gaze direction, fixed stares, or inconsistent eye tracking due to animation errors.

Head pose estimation: Estimating the 3D orientation of the head can reveal mismatches between facial movement and head motion. In deepfakes, facial features might not align properly with the head's rotation or position.

Facial micro-expressions: Micro-expressions are involuntary facial movements that reflect emotions. Deepfake generators can often fail to replicate these subtle expressions, leading to emotionally flat or robotic faces.

Erratic movement: Deepfakes may show unnatural or jerky motion, especially during rapid head turns or body movements. This results from inconsistencies in frame synthesis or poor temporal coherence.

Mismatching features: Facial features such as the ears, teeth, or jawline may appear misaligned or inconsistent across frames, especially under different lighting or angles.

Absence of fine detail ("airbrushed" look): Skin texture, pores, or subtle facial imperfections may be smoothed over or missing entirely, giving the face an overly polished or artificial appearance.

Nonhuman features (e.g. extra fingers or deformed ears): AI-generated images or video frames may occasionally produce anatomical errors such as too many fingers, misshapen ears, or asymmetrical limbs, particularly in low-quality or quickly produced deepfakes.

j) Software inspection and code-level analysis

In certain forensic scenarios, particularly when dealing with synthetic or manipulated media, it may be necessary to inspect the software tools suspected of generating the content. This involves examining the source code or reverse-engineering the application to identify how it processes input, generates output, or leaves identifiable artefacts. Code-level inspection can reveal default parameters, rendering methods, compression routines, or naming conventions that may leave behind subtle but consistent traces in the resulting media. These artefacts, whether in file structure, metadata, or signal characteristics can help attribute content to specific tools or workflows, strengthening the evidentiary value of the analysis. While not always accessible or feasible, software inspection is a valuable line of inquiry when source tools are known or suspected.

4.3 Audio[35]

a) File format analysis

Linking with its use for provenance of material, file format analysis establishes whether the purported suspect file format is in keeping with the contents of the material, essentially asking where the purported device have created the file. There is evidence to suggest that certain elements of content will be missing in the case of synthetic material. In addition, there is evidence to suggest some deepfake and synthetic material generators may in fact use a proprietary version of open-source formats, with unique levels of content that may not necessarily align with the original design of the file containers.

b) File structure analysis

File structure analysis is a method of identifying if the file is device-original or if there are any opportunities for manipulation. Magnet Verify (formerly Medex) is one example of a tool designed to compare the metadata of a file with the expected structure of that file type at the binary level. Analysis of matching/mismatching containers and internal organization of the data comprising the file itself can reveal indications of the provenance of the material.

c) Quantization-level/Bit-depth analysis

A forensic method that examines the resolution at which an audio signal has been digitally encoded is called quantization-level or bit-depth analysis. Bit depth determines the number of possible amplitude values available to represent the signal, directly affecting dynamic range and noise floor. Inconsistencies in quantization levels may reveal transcoding, editing, or re-recording, making this analysis valuable for detecting manipulation and assessing the authenticity of digital audio. This type of analysis can only be

[35]ENFSI et al. (2022) Best Practice Manual for Digital Audio Authenticity Analysis, ENFSI-FSA-BPM-002, pp.2–26. Available at: https://enfsi.eu/wp-content/uploads/2022/12/FSA-BPM-002_BPM-for-Digital-Audio-Authenticity-Analysis.pdf.

conducted if the recorder model, recording software and version, and the claimed recording settings are known, and if the submitted file is provided in an uncompressed Pulse Code Modulation (PCM) format.

d) DC offset

Specifically in audio, the waveform normally oscillates symmetrically above and below the 0-line, thus creating the amplitude. In case of offset, this 0-line is shifted up or down (+/-). Device fingerprinting, editing, detection, and tampering can be identified using this investigative method. DC-offset analysis should not be used to attribute a recording to a specific device, but rather to eliminate certain devices from the pool of candidates.

e) Long-term spectral analyses (LTSA)

Spectral analysis is breaking an audio signal into its constituent parts and displaying them in a way that each can be seen at the same time. Normally, this is completed via spectrographic visualization as a result of a fast Fourier transform (FFT). This visualization is key to completing many detection methods, though understanding what is represented by this transform is key. Identifiers such as inconsistencies in background noise, the noise floor shifting abruptly, and patching together separate sections that have been "generated" can indicate the presence of synthetic media. Real speech normally has consistent background noise; however, synthetic voices generally lack these markers, which can be identified through LTSA.

f) Compression-level analysis

Generated video, image, and audio files can be re-encoded differently than genuine device-created media. Compression-level analysis is a forensic technique used to examine the degree of lossy or lossless compression applied to digital media files, such as audio, video, or images, to detect editing, tampering, source authenticity, or processing history. Different compression levels can leave behind identifiable artefacts that help determine whether a file has been altered or re-encoded.

g) Modified discrete cosine transform analysis (MDCTA)

This is a mathematical technique used to present audio signals in the frequency domain rather than as raw waveforms. It is a core component of many audio compression algorithms such as MP3 and AAC. It breaks the signal into overlapping time "windows" making subtle changes in frequency content visible. Synthetic speech often exhibits unnatural frequency patterns or anomalies across different "windows". Real human speech has smooth and complex spectral transitions, whereas generated audio may show signs of over-smoothing, spectral leakage, or blocky artefacts appearing as a result of its synthesis or compression. MDCTA is another method of looking at these features. It can also complement time-domain (waveform) or spectrogram-based methods.

h) Electric network frequency (ENF) analysis

ENF analysis is an authentication technique that examines the fluctuations of mains electricity frequency (50 or 60 Hz) that can be unintentionally captured in audio and video recordings. These fluctuations act as a unique "time and location stamp", allowing investigators to determine when and sometimes where a recording was made.

i) Discontinuity analysis in ENF

Discontinuity analysis in ENF examination involves detecting abrupt breaks, jumps, or misalignments in the frequency trace embedded within an audio or video recording. Since ENF naturally varies smoothly over time, sudden discontinuities may indicate editing events such as splicing, deletion, or reordering of segments. By comparing the captured ENF pattern against a reference database or expected continuity, investigators can identify points where the recording has likely been manipulated.

j) Frequency response analysis

This audio forensic technique is used to examine segments of a recording, particularly where speech predominates, to evaluate the unique acoustic characteristics of the recording device. By analysing the frequency spectrum of the audio, it can be compared against known profiles of candidate microphones or recording devices to determine which is most likely to have been used to produce the recording. In the case of synthetic media detection, the absence of such indicators of the microphone may raise suspicion about the origin of the file. In addition, unnatural responses may indicate the presence of synthetic generation.

k) Microphone thermal noise analysis

This analysis examines the inherent electrical noise generated by a microphone's components due to thermal agitation of electrons. By characterizing the spectral and amplitude properties of this noise, investigators can assess the microphone's identity, quality, or condition, and in some cases, detect anomalies indicative of tampering, modification, or device inconsistencies.

l) Statistical encoding analysis

Statistical encoding analysis examines the patterns and distributions of data within a digital audio file, such as bit patterns, sample values, or compression artefacts, to determine characteristics of the recording process or file format. It can help identify anomalies, inconsistencies, or manipulations introduced during recording, editing, or compression, and may provide insight into the software or device used to produce the audio.

m) Inverse decoding

Inverse decoding is a forensic technique used in PCM files to reverse the effects of audio compression or encoding, aiming to reconstruct the original signal as closely as possible. By analysing the encoded file and applying knowledge of the compression algorithm, this method can reveal subtle artefacts, distortions, or evidence of manipulation that may not be apparent in the compressed recording.

n) Copy-move forgery detection

Copy-move forgery occurs when a segment of an audio recording is duplicated and inserted elsewhere within the same file to manipulate content. Detecting such forgeries typically involves the following methods:

i. Correlation analysis – Comparing short segments of the waveform against other parts of the recording to identify identical or highly similar content. High correlation between non-adjacent segments may suggest duplication.

ii. Signal subtraction – Subtracting suspected repeated segments from the recording. If the segment cancels out or leaves minimal residual energy, it may confirm a potential copied portion.
iii. Audio fingerprinting – Generating unique fingerprints for small audio segments and searching for repeated fingerprints within the same file to locate duplicated material.
iv. Spectrographic analysis – Visualizing the frequency content of the recording. Repeated segments often show identical spectral patterns, making them easier to spot visually.
v. Auditory verification – Listening to the flagged segments to confirm the automated detection, ensuring that context or natural repetitions are not misinterpreted as forgeries.

o) Waveform analysis

Using the signal in its classic representation of the waveform, the direct amplitude signal over time, subtle markers emerge that may initially appear unnatural. Features include sudden increase or decrease in amplitude, repeating and unexplained patterns surrounding speech features, or phase irregularities within the waveform, since generative models generally prioritize perceptual quality over physical accuracy.

p) Signal power

A measure of the energy or intensity of an audio signal over a given period. Variations in signal power can reveal inconsistencies, sudden level changes, or unnatural transitions that may be linked to editing or tampering.

q) Spectrographic analysis

A visual representation of an audio signal that displays frequency, time, and intensity simultaneously. Spectrograms make it possible to detect patterns, background noises, or frequency anomalies that may be overlooked in waveform-only analysis, providing a more detailed forensic view of the recording.

r) Cross-fade identification

Some synthetic media generation models create individual sounds that are then stitched together to create a cohesive file – known as a composite or compilation file. At the points where the individual sounds are stitched together, some models use a “cross-fade” to cover these points and restrict “pop” noises at the point of the file cross over – smoothing over the gaps where these foundational files meet by “fading out” the preceding clip, and “fading in” the proceeding. These cross-fades, often found through spectrographic analysis, may also be an indication of intentional editing, as it is a basic technique with audio engineering and music production used widely across the media industry. Conversely, the absence of these cross-fades may have other implications that indicate the presence of synthetic media or manipulation.

s) Zero-amplitude padding

In forensic audio, zero-amplitude padding refers to stretches of “perfect digital silence” (no waveform activity at all) that are inserted into a recording. Unlike natural pauses, which always contain some background noise, mic hiss, or room ambience, artificially padded silences are completely flat. These gaps often appear when audio has been cut, spliced, or generated by algorithms, making them a useful indicator of potential manipulation or synthetic tampering.

t) Butt-splice detection

A forensic technique used to identify points where two audio segments have been directly joined together. Detecting such splices can provide evidence of editing, omission, or reordering of speech or events within a recording.

u) Generation of sample recordings/exemplars

The creation of controlled recordings for comparison with questioned material. Exemplars can be used to evaluate voice identity, equipment characteristics, or acoustic conditions, offering a reliable baseline for forensic comparison and authentication.

v) Speaker comparison

Linguistics and speech analysis is already an established forensic science with many specialists in the field worldwide. For some time, this has been the main focus of audio forensics as a discipline; however, this approach is now changing. Research has shown that using these techniques may indeed show evidence of synthetic media being present within speech patterns. By analysing the frequency structure of a voice, there are several possible indicators that may prove the presence of deepfake elements. Normally, to conduct this, a known sample of the suspected voice is required for the suspicious sample to be compared against. This is not always possible within an investigation. For deepfakes, it may be possible; however, for entirely synthetic voices, this is not possible.

w) Audible anomalies

Quite often, the audible anomalies are also visual when looked at within spectrographic analysis. The presence of these features can make synthetic voices appear robotic or "too perfect." However, the identification of these features can be difficult due to their presence being explained by several reasons other than as a result of synthetic creation.

- Unnatural prosody – rhythm, stress, and intonation that do not match human pattern or capability.
- Pronunciation and enunciation – the way in which the words and phrases are said correctly, in addition to the clarity with which they are spoken so that others may understand.
- Glitches/artefacts – clicks, pops, or robotic tones that do not move or flex, introduced by imperfect generation.
- Breath and pause irregularities – speech without surrounding realism can indicate a lack of humanity during creation and/or capture. May be resultant of training data being too "clean", where added data or noise may have been removed.
- Phase or timbre distortions – sometimes referred to as the "colour" of the voice sounding hollow, metallic, overly smooth, or granular.
- Inconsistent background noise – shifts in noise floor that do not align with natural recording from purported device.

x) Audible artefacts

Audible artefacts are unnatural or unintended sounds introduced during recording, compression, manipulation, or synthesis of audio. These may include clicks, pops, distortions, digital glitches, unnatural pitch shifts, or inconsistent background noise. In the context of synthetic or edited audio (such as voice deepfakes), artefacts can signal tampering, poor synthesis quality, or traces of underlying manipulation.

y) Contextual analysis

This involves evaluating an audio recording not just by its technical properties, but by examining the content, setting, and circumstances surrounding it. This includes assessing whether the dialogue makes logical sense, if background sounds match the claimed environment, or whether speaker behaviour aligns with expectations. In forensic work, this helps identify inconsistencies that may suggest deception, fabrication, or manipulation, especially when combined with technical findings.

z) Reverberation inconsistencies

The unnatural or abrupt changes in the ambient echo or room acoustics within a single recording are called “reverberation inconsistencies.” Reverberation is shaped by the physical environment in which audio is captured – its size, materials, and layout. If the reverberation profile shifts unexpectedly (e.g. from dry to echoey, or from one spatial signature to another), it may indicate that segments were recorded in different locations, captured using different equipment, or artificially inserted. These inconsistencies are often subtle and require detailed acoustic analysis to detect, but they can serve as strong indicators of manipulation or fabrication.

5. AI AND MACHINE LEARNING

There is currently a major debate within the audiovisual forensic community as to whether AI and ML tools should be used within evidential processes. There are several arguments for and against; however, one must consider the risk to the individual items of evidence.

5.1 Explainability

The current explainability of deepfake detection models powered by AI and ML are almost non-existent. Most proprietary "detectors" are "black box" systems, of which it is impossible to determine by which factors (loci) a system deems a file to be a positive or negative match for deepfake content.

Of the forensic tools currently available that utilize this technology, a percentage confidence is displayed after the suspect file has been processed to represent the result of the analysis. In traditional forensics, for example DNA analysis, for two samples to be considered a positive match, there must be a minimum of 95 per cent confidence for it to be defensible in court. In this case, we can identify the exact loci by which this has come to be; therefore, it is explainable and fully repeatable by peer review. This process is also guaranteed to be standardized and conducted the same way across all DNA cases. However, in the context of AI detection of deepfakes, the black box nature of the systems means we cannot identify the loci, nor can we explain how the system has come to its percentage confidence conclusion. This means that the result of the system's analysis is, in fact, a probabilistic guess, rather than a validated statistical conclusion. Until such time as these systems can be fully explainable, it is difficult to justify their use in forensic science, especially considering the lack of proven manual methods by which to check and authenticate a model's detection. There is not yet a place in forensics where these systems can be used for processing unsupervised and unchecked.[36]

5.2 Intelligence uses

Given that the forensic uses have significant limitations and caveats due to the priorities and methodology, there is a case for using these tools for intelligence purposes within an investigative/intelligence setting. When addressing forensic uses, it is assumed that these files will then be put forward for evidence within legal proceedings. However, the use of these AI and ML algorithms may come into its own when attempting to identify a line of inquiry at the policing end of the legal system, not necessarily within forensic processing.

5.3 Intelligence-to-evidence approach

Should AI or ML tools be introduced into the forensic workflow, there is currently little possibility of them being used without human oversight, which will likely be the case for the foreseeable future. However, using these methods for investigative purposes (investigative approach), of which the result is then verified by the human practitioner (forensic approach), still requires the following:

[36] Jackson, A.R.W. and Jackson, J.M. (2017) *Forensic science*. 4th edn. Harlow: Pearson Education Limited.

- Fully explainable and transparent loci within the neural network.
- Manual methods by which to verify these loci.
- Practitioners with suitable expertise by which to verify the loci.

If an AI or ML system is the best option available to a unit, then it should not be used in isolation. There is currently no eventuality where an AI would be allowed to function "unsupervised" or unchecked. The need for human intervention and verification of the results is still very real. Until sufficient functionality and explainability of these tools is achieved, trust should not be placed in the tool alone.

5.4 Occam's razor

Also known as the Law of Parsimony, the principle gives precedence to simplicity: of two competing theories, the simpler explanation of an entity is to be preferred.[37] In essence, the solution containing fewer assumptions is the preferred course of action.

In the case of involving AI and ML within digital forensics, there are significant assumptions to be made, including the assumption that the decision loci are correct – adding complexity without clarity and layers of abstraction. Whereas with human oversight and explainable tools, the number of assumptions is fewer and therefore more viable, not just within a forensic context.

In forensic science, methods should not multiply complexity without necessity. AI-based systems may offer potential, but unless their benefits clearly outweigh the added opacity of their processing, they should be approached cautiously.

5.5 Validation

Since policing and forensics are incredibly sensitive workflows, the tools chosen for detecting synthetic media must be subject to rigorous validation in line with policies and legislation of the relevant institution and jurisdiction. Validation is not simply a technical formality, it ensures reliability, accountability, and trust in both forensic and non-forensic settings. At a minimum, it ensures that due diligence has been undertaken in the careful selection of tools being considered for inclusion on these workflows.

When considering new tools, the following questions are applicable in this sense:

1. Independent validation
   Has the method been tested and reviewed by organizations beyond its developer?
2. Real-world robustness
   Does it perform reliably on degraded, compressed, or deliberately manipulated media – not just in controlled environments?
3. Transparency and explainability
   Does the method provide interpretable results that can be explained at a functional level?
4. Repeatability and reliability

[37] *Occam's Razor* (2025) *Encyclopædia Britannica*. Available at: https://www.britannica.com/topic/Occams-razor. Last accessed on 11 August 2025.

Can the method produce consistent outcomes across users, systems, and contexts? Is it resistant to small changes in input?

5. Integration into workflow
   Can it be embedded within existing investigative and chain-of-custody processes without compromising evidence handling or operational security?
6. Lifecycle and updating
   Is there a process for ongoing updates and re-validation to ensure resilience against rapidly evolving generative technologies?
7. Governance and oversight
   Are there mechanisms for accountability to prevent unverified or "snake oil" tools from being introduced into critical workflows?

These questions are not just relevant in the validation phase if introducing a new tool into a new organization, but they also ask another question: is it forensically viable? This may not be required in policing, but once used in a forensic setting, this becomes an extremely important decision point, else it cannot be used.

Model efficacy

Detection model performance can potentially drop significantly when trained on clean data designed with the intention of being used as part of that training dataset, compared to when being used on real-world, noisy, or tampered content. As it stands currently, real media are being flagged as fake, purely because they are of poor quality, possibly due to the loci of detection being obscured, among other reasons. It again highlights the importance of balanced training data that do not simply utilize worldwide and open-source data sets.

5.6 Deep learning for detecting deepfakes

Some of the most advanced tools for spotting deepfakes use AI itself. These tools rely on a type of AI called "deep learning", which means training computer systems to recognize patterns in images and videos, especially patterns that humans might miss.

There are different kinds of deep learning models that help with this:

CNNs (convolutional neural networks) are excellent at looking at single images or video frames. They can spot strange textures, overly smooth skin, or tiny details that do not look quite right.

RNNs (recurrent neural networks) are used to understand how things change over time, like how a face moves in a video. Two special types of RNNs are **LSTMs (long short-term memory networks)** and **GRUs (gated recurrent units)**. These models are designed to remember patterns across a sequence of frames. For example, they can catch if a person's face shifts slightly in an unrealistic way, or if blinking and expressions do not flow naturally.

Vision transformers are a newer type of model that look at an image in pieces and then figure out how all those pieces fit together. This helps them find more complex or widespread deepfake clues.

Multimodal models combine both video and audio inputs. They can detect if the sound does not match the person's lips or if the voice sounds too robotic or artificial. Models like CNNs, RNNs, vision transformers, and multimodal networks can detect artefacts such as over-smoothed skin, warping, irregular blinking, poor lip-sync, or unnatural motion. These approaches learn subtle inconsistencies that are difficult for traditional detection methods to catch.

These deep learning systems are trained on huge collections of real and fake videos and images. Over time, they learn what looks natural and what looks off. In some cases, they can be better than older tools at finding subtle signs of fakery and can help us tell whether the media are real or AI generated.

However, it is also important to understand that using AI to catch AI can raise challenges, especially in legal or forensic contexts. Deep learning models are often considered "black boxes," meaning we may not fully understand how they arrive at their conclusions. This lack of transparency and "repeatability" can make it difficult to use their results as solid evidence in court or formal investigations. Even so, these tools are valuable for raising red flags and should be recognized for their awareness value and potential early detection purposes.

6. TRAINING AND EDUCATION IN DEEPFAKE DETECTION

Considering that a large amount of current forensic detection capabilities are highly technical, qualified individuals are essential to conduct these methodologies, especially when it is required for them to defend their work.

The landscape of digital forensics training appears to be fragmented into three sectors:

(a) Academic

Through validated and accredited universities and educational centres, often carrying academic credits.

Pursuing higher-level training at a recognized academic institution with formal certification and qualification is the more traditional approach of education. However, this process is slow to complete, and often the content does not change as often as the industry. Moreover, there is usually limited applicability to the job role(s) it is designed for.

(b) Vocational

The vocational training path could be taken via policing institutions; this route is designed to be more contextualized than the academic and could be considered as an add-on to the academic qualifications, with the extra benefit of learning while on the job, so it can be taken back to the role immediately.

The vocational training mode enables pragmatism within the modules of the course. This adaptability means that new knowledge can be disseminated among those who take part in the training with a faster turnaround than other training modes. This is a highly valuable tool, considering the landscape of synthetic media is changing day to day

(c) Proprietary

Software developers mainly train on how to use their products. Known as “vendor-specific training,” it ensures that the practitioner is able to prove they are using the software in the correct manner, especially if it is a highly nuanced and advanced software.

These vendors, through market competition, therefore help drive the development and advancement of practices in the field, often feeding new and cutting-edge information and techniques to the practitioners taking advantage of the software and training available. However, since the training is software-specific, the other aspects of the practitioner role may be overlooked and not present a true reflection of the job in which the software may be used.

Each area of training has its own strengths and weaknesses. In the pursuit of accredited and validated training, collaborative efforts are by far the most appropriate approach. Furthermore, the development of detection methodologies must be a collaborative approach between all three sectors, due to the same strengths and weaknesses in the training abilities. INTERPOL is in a position to encourage this collaboration across these sectors, so that training reflects both scientific rigour and operational realities. This fosters well-rounded practitioners who understand the merits of both mentalities, reducing the siloed

expertise that is currently exacerbating the issue surrounding deepfake detection within and without law enforcement.

6.1 Ground truth dependency

Authentication of a file requires comparing the questioned item to an established standard or ground truth. Without such a reference, the conclusion is not scientifically justifiable. This ground truth dependency is standard forensic practice and should be considered at the point of authentication of suspect deepfake files. The principle of ground truth dependency in forensic science makes adequate training essential in every media domain. Whether analysing audio, video or still images, analysts must be able to recognize authentic patterns before they can justify identifying fakes. That expertise is built only through rigorous and domain-specific training. Each medium demands the same level of respect and importance, due to the high likelihood of an error in one domain translating into another.

6.2 Impostor bias

Insufficient training in media forensics in general may lead to the appearance of impostor bias.[38] This behaviour occurs due to the awareness of AI-generated content. Practitioners and those examining suspect specimens may systematically doubt the authenticity of legitimate audio, videos, and images, despite them being or appearing to be genuine. This is not based on the content of the files themselves, but on preconceived assumptions about their origin. Without adequate training in these media fields, it is not possible to combat this bias. The provenance of the material must be established in accordance with forensic science principles, but must also be considered when the question of synthetic media authentication comes to light. This issue has been recognized in contemporary research, but is a growing threat to the integrity of the forensic investigative processes, without a proper baseline understanding of genuine media.

6.3 Standardization and accreditation

The training programmes currently available vary widely in scope, quality, and recognition. Accreditation bodies such as regional policing organizations, academic boards, or international forensic associations should establish baseline competencies for those involved in synthetic media detection, especially within the forensic realm. Certification of a practitioner in this area should not only focus on technical competence and detection skills, but must also consider evidentiary standards, ethics, and cross-media awareness. Training programmes should align with shared principles of transparency, reproducibility, and practitioner competence.

6.4 Critical tool literacy

[38] Casu, M., Guarnera, L., Caponnetto, P., & Battiato, S. (2024). GenAI mirage: The impostor bias and the deepfake detection challenge in the era of artificial illusions. In *Forensic Science International: Digital Investigation* (Vol. 50). Elsevier Ltd. https://doi.org/10.1016/j.fsidi.2024.301795

Training should not only cover how to use the detection tools and methodologies, but also how to evaluate their validity, limitations, and margins of error. An over-reliance on "push-button" solutions will inevitably raise questions about the validity of that practice. As in forensic science theory, those processing and analysing the evidence should be knowledgeable on the effects of said process on the file itself – does it alter the colour values, cause loss of information, remove original data, and questions alike. Knowing the pros and cons of each process is also reflected in the tool literacy skills of the practitioner. Without knowing the tools and how they function, along with any possible adverse effects, an unknown adverse effect is likely imposed on the evidence, and possibly later affecting the outcomes of the assessment.

6.5 Courtroom readiness

Robust training and theoretical knowledge will ensure that practitioners and experts are prepared to explain their workings and methodologies clearly, withstanding any adversarial scrutiny in judicial proceedings. The threat of synthetic media appearing in forensic evidence has shone a light on the current practice of detection, especially within policing, and with the threat of the liar's dividend being used as a credible defence within court proceedings, it is no longer acceptable to rely on that gut instinct – this does not stand up to scrutiny. We need to defend our decisions appropriately to avoid that risk becoming reality.

Courtroom readiness for defending synthetic media detection, in line with forensic science theory, requires three things:

1. The ability to articulate the methods used, their limitations, and the risk posed to the evidence by undertaking those methods.
2. Ensuring conclusions are repeatable, defensible, and backed by appropriate validation.
3. Preparing experts to give testimony under adversarial questioning, where the credibility of both the processed evidence and the experts themselves may be challenged.

By embedding these principles into the practice of deepfake detection in policing, we can ensure that the credibility of the practitioner and the institution, by extension, is maintained and that the media analysis is accepted as robust in judicial settings.

6.6 Continuous professional development (CPD)

The great AI evolution is continuing and will continue moving regardless of the developments that are made. As new solutions are brought into play, these should be integrated into policing through continuous professional development. Without CPD, old techniques may prove ineffective or no longer applicable. The learning should be ongoing, not a one-time training set that will solve the issue once and for all. One way that INTERPOL is already supporting its community of practitioners on the options available to tackle the issue is by coordinating updates on emerging trends and techniques. This is an approach that should become more commonplace, ensuring that practitioners remain current with evolving threats and maintain credibility in investigative and judicial contexts. This mechanism also supports the testing of these new approaches as they become available, ensuring that they are practicable, not just theoretical, so that they can be deployed under the correct conditions.

7. REGULATION AND LEGAL CONSIDERATION

Despite growing societal awareness of deepfakes and shallowfakes, most jurisdictions still lack clear and targeted legislation to address them. Law enforcement agencies are often left relying on outdated statutes, such as defamation, harassment, copyright infringement, and other unsuitable categories, which are ill-suited to deal with the speed, scale, and anonymity of synthetic media threats. This has resulted in a patchwork of legal frameworks that vary widely between countries, making consistent enforcement and cross-border cooperation especially challenging. For policing agencies, such legal fragmentation complicates investigations and can leave victims without a clear path to justice.

A recent legislative proposal from Denmark[39] illustrates a promising approach. In response to high-profile deepfake incidents, the Danish government has since proposed legislation that would grant individuals legal ownership over their face, voice, and likeness. Under this proposed model, unauthorized AI-generated content impersonating an individual could be subject to removal and damages claims. While the law is designed for civil enforcement and stops short of criminalizing the creation and distribution of deepfakes, it is a first step, and it does introduce meaningful tools for victims and obliges platforms to act under the EU's Digital Services Act.[40] This model still relies on platforms and civil litigation, and criminal accountability remains elusive in most cases.

Deepfakes and shallowfakes present a unique evidentiary challenge for law enforcement: the liar's dividend. This refers to the strategic denial of authentic digital evidence on the grounds that it may be AI-generated. As synthetic content becomes more realistic, suspects or political actors may exploit uncertainty around the authenticity of digital materials to try and discredit them. This not only complicates investigations and prosecutions, but also undermines trust in digital evidence presented in court. Currently, many agencies do not possess the specialized capabilities needed to identify and verify synthetic media. In most jurisdictions, analysis of images, audio, and videos will fall under general forensics and digital forensics units, which may lack the sector-specific expertise to detect deepfakes. Without dedicated audiovisual forensic support, investigations may be delayed or even dropped due to doubts in reliability. There is a clear need for agencies and forensic laboratories to invest in audiovisual specialization.

The international guidelines below may be taken into account when designing a workflow and standards in adopting the practice of synthetic media detection.

Use of AI in forensics:

ISO/IEC 17025[41] – General requirements for the competence of testing and calibration laboratories.
ISO/IEC 17020[42] – Requirements for inspection bodies. Ensuring methodologies are conducted impartially and competently.

[39] Woods, C. (2025) Denmark proposes copyright laws to protect against deepfakes - Law Society Journal. https://lsj.com.au/articles/denmark-proposes-copyright-laws-to-protect-against-deepfakes/. last accessed on 25 June 2025.
[40] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), *Official Journal of the European Union*, L 277, 27 October 2022, pp. 1–102, https://eur-lex.europa.eu/eli/reg/2022/2065/oj., last accessed on 25 June 2025.
[41] ISO/IEC. (2017) *General requirements for the competence of testing and calibration laboratories* (ISO/IEC 17025:2017). https://www.iso.org/standard/66912.html, last accessed on 13 February 2026.
[42] ISO/IEC. (2012) *Conformity assessment — Requirements for the operation of various types of bodies performing inspection* (ISO/IEC 17020:2012). https://www.iso.org/standard/52994.html, last accessed on 14 July 2025.

ISO/IEC 21043[43] – Specifically for forensic science, requirements and recommendations ensuring quality of process.
ISO/IEC 27041[44] – Guidance on adequacy of incident investigation methods.
ISO/IEC 27042[45] – Analysis and interpretation of digital evidence.
ISO/IEC 42001[46] – Management system standards specifically for artificial intelligence.
European Network of Forensic Science Institutes (ENFSI) Best Practice Manuals[47] [48] [49] – European forensic guidelines. While region-specific, they inform expectations around incorporating emerging technologies into practice.
Scientific Working Group on Digital Evidence (SWGDE) Best Practices[50] [51] [52] – While again region-specific to the United States, this guidance on handling digital evidence is frequently cited internationally, offering baseline practices for digital forensics.

Deepfake detection & synthetic media provenance:

ISO/IEC 27037[53] – Guidelines for collection and preservation of digital evidence which present globally applicable methods for identifying, collecting, and preserving digital evidence, including synthetic media.
ISO/IEC 27043[54] – Incident investigation principles. Critical when probing the provenance of manipulated or deepfake content.
C2PA[29] – Content provenance and authenticity standard. An open standard for embedding cryptographic provenance into media files. Enables the verification of origin and establishes integrity.
NIST SP 800-86[55] – A guide to integrating forensic techniques into incident response. Internationally respected guide offering comprehensive forensic workflow design. This is highly applicable to deepfake investigations where response speed and method rigour are essential.

[43] ISO/IEC. (2025) *Forensic sciences — Part 3: Analysis* (ISO/IEC 21043-3:2025). https://www.iso.org/standard/73896.html, last accessed on 11 August 2025.
[44] ISO/IEC. (2015) *Information technology — Security techniques — Incident investigation principles and processes* (ISO/IEC 27043:2015). https://www.iso.org/standard/44405.html, last accessed on 14 August 2025.
[45] ISO/IEC. (2015) *Information technology — Security techniques — Guidelines for the identification, collection, acquisition and preservation of digital evidence* (ISO/IEC 27037:2015). https://www.iso.org/standard/44406.html, last accessed on 22 July 2025.
[46] ISO/IEC. (2023) *Information technology — Artificial intelligence — Management system* (ISO/IEC 42001:2023). https://www.iso.org/standard/42001.html, last accessed on 18 June 2025.
[47] ENFSI (2018). *Best Practice Manual for Forensic Image and Video Enhancement* (Version 01). Available at: https://enfsi.eu/wp-content/uploads/2017/06/Best-Practice-Manual-for-Forensic-Image-and-Video-Enhancement.pdf, last accessed 14 September 2025.
[48] ENFSI. (2021) *Best Practice Manual for Forensic Image Authentication* (ENFSI-BPM-DI-03). https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf, last accessed on 14 August 2025.
[49] ENFSI. (2022) *Best Practice Manual for Forensic Digital Audio Authenticity Analysis* (FSA-BPM-002). https://enfsi.eu/wp-content/uploads/2026/01/FSA-BPM-002_Digital-Audio-Authenticity-Analysis-FIXED.pdf, last accessed on 12 August 2025.
[50] SWGDE. (2022) *SWGDE Best Practices for Forensic Audio* (Version 2.5). https://www.swgde.org/wp-content/uploads/2023/11/2022-06-09-SWGDE-Best-Practices-for-Forensic-Audio_v2.5.pdf. last accessed on 14 August 2025.
[51] https://www.swgde.org/wp-content/uploads/2025/03/2025-03-03-Best-Practices-for-Image-Authentication-18-I-001-2.0.pdf
[52] SWGDE. (2025) *Best Practices for Digital Forensic Video Analysis* (Version 2.0). https://swgde.org/wp-content/uploads/2025/12/2025-12-10-Best-Practices-for-Digital-Forensic-Video-Analysis-18-V-001-2.0.pdf, last accessed on 19 August 2025.
[53] ISO/IEC. (2012) *Information technology — Security techniques — Guidelines for identification, collection, acquisition and preservation of digital evidence* (ISO/IEC 27037:2012). https://www.iso.org/standard/44381.html. last accessed on 15/06/2025.
[54] ISO/IEC. (2015) *Information technology — Security techniques — Storage security* (ISO/IEC 27040:2015). https://www.iso.org/standard/44407.html. last accessed on 03 August 2025.
[55] Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Special Publication 800-86 Guide to Integrating Forensic Techniques into Incident Response Recommendations of the National Institute of Standards and Technology*. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-86.pdf., last accessed on 15 August 2025.

8. CONCLUSION AND RECOMMENDATIONS

The landscape of this issue is changing daily – not only are cases beginning to be affected by the use of synthetic media, but the rhetoric surrounding them is now being used to raise doubt about the authenticity of real media. There appears to be a rise in "snake oil" solutions to tackle the existence of synthetic media, not just in forensic science but also in the wider world. Meanwhile, the development of generative tools is becoming much more sophisticated and increasingly difficult to detect. Yet the answer to detection does not necessarily rely on AI or ML – it is entirely dependent on the use case.

AI-manipulated media are no longer a future threat; they are here, pervasive, and increasingly sophisticated. Deepfakes and other AI-altered content can infiltrate investigations silently, blending with authentic evidence in ways that evade traditional detection. It is crucial to ask ourselves if we are truly certain that our chain of custody is free from AI interference. Without targeted measures, AI-manipulated content may already be present, undetected.

The solution starts with AI media literacy for law enforcement. Investigators must understand how AI manipulates audio, images, and videos, how to verify authenticity, and which forensic tools are reliable. AI literacy is no longer optional; it is essential for maintaining the integrity of evidence.

Given the rapid pace of AI development**,** continuous research collaboration and international knowledge sharing are critical. Techniques that are state of the art today may be obsolete tomorrow. Beyond technical innovation, there is a clear need to consider research protections within countries and at higher security levels. Introducing and enforcing policies is vital – the dissemination of deepfakes is not purely a technological problem, it is in the hands of people as well as private companies, and all actors share responsibility in mitigating risk. Bias in perception further complicates detection. AI manipulations exploit assumptions and human error, making multi-layer forensic approaches essential.

This document is not static; its purpose is to be a living, evolving resource. The AI landscape moves so quickly that research and documentation can lag behind real-world threats. Regular updates and revisions are required to keep guidance current and actionable. Ultimately, AI knows no borders, respects no rules, and challenges every traditional notion of authenticity. Maintaining confidence in digital evidence requires proactive, informed, and collaborative action: ongoing education, adaptive frameworks, international cooperation, and commitment to robust detection strategies.

After reading this, the question remains: can we confidently say our evidence is free from AI interference? If the answer is anything less than a resounding yes, it is time to act – develop literacy, implement frameworks, enforce protective policies, and stay vigilant. The integrity of evidence depends on it.

To mitigate and address the current threat within the individual justice systems and cross-border investigations, the following recommendations have been suggested.

Short term

Develop AI media literacy for law enforcement.

Officers, investigators, and forensic specialists should have a functional understanding of how AI systems work and how it is deployed in their units, if at all. This includes basic technical knowledge, awareness of legal boundaries, and understanding possible ways AI manipulation may appear in different cases. Training should be aligned with national legislation and local law enforcement policies to avoid misuse.

Promote shared responsibility between individuals and tech companies.
Detection and prevention of deepfakes must be a joint effort. Law enforcement should work with technology providers to ensure solutions are practical and not just theoretical. This includes securing access to relevant test material, establishing evaluation criteria, and ensuring adherence to professional and ethical policy frameworks.

Address bias in the detection of videos, images, and audio.
Current forensic practice tends to prioritize visual media, creating a gap in audio expertise. Upskilling in audio capabilities is needed to ensure parity across all modalities. Recognizing audio processing as a dedicated forensic discipline will help detection methods in this area reach the same maturity as video and image detection.

Introduce training initiatives for video, image, and audio handling and forensics.
Training should cover both the handling of authentic media and the identification of the manipulated content. Practitioners must first establish strong expertise in working with unaltered media before advancing to detecting synthetic or manipulated content. This builds resilience, ensuring that training addresses the fundamentals as well as advanced detection.

Medium term

Provide regular training and education programmes.
Since the landscape of this threat is changing, so should the knowledge available. Training programmes must therefore be recurring and quick to update, ensuring officers and forensic specialists can stay informed about the latest manipulation techniques, tools, and countermeasures.

Ensure continuous development of new detection techniques.
Law enforcement should work with academics and industry alike to encourage iterative development of detection methods. Tools should be benchmarked, validated, and stress-tested across diverse datasets to reflect real-world conditions.

Continuous review of developing risks to security and justice.
Deepfake risks may shift between media types – sometimes audio is the biggest threat, at other times videos, or images. Ongoing risk assessments are required to ensure responses remain flexible and proportional to emerging risks.

Long term

Build adaptive international frameworks for AI threats.
Since generative AI tools and manipulations cross borders, international collaboration is essential. Establishing adaptable frameworks will allow law enforcement agencies to share expertise, intelligence, and detection resources across jurisdictions.

Establish research protections at political and security levels.
Research into detection should not remain solely in academia, where publication pressures can sometimes expose vulnerabilities and detection opportunities within generative systems. Mechanisms should be created for sensitive detection methods to be securely shared with law enforcement and relevant security institutions, while restricting broader release when risks outweigh the benefits.

Introduce national and international legislation.
Legal frameworks should clearly define the misuse of deepfakes, providing investigative powers for law enforcement and setting standards for accountability. International harmonization of legislation will strengthen collective resilience.

Foster ongoing research collaborations.
Strong, long-term partnerships between law enforcement, industry, and academia are absolutely vital. Shared research centres, joint projects, and co-funded initiatives can ensure innovation continues and that policing agencies remain ahead of evolving threats.

GLOSSARY OF TERMS

Artificial Intelligence (AI)
Computer systems designed to perform tasks that typically require human intelligence, such as pattern recognition, decision-making, or language processing.

Attribution
Identifying the source or actor responsible for creating or disseminating manipulated content.

Audio
Sound-based digital media, including recordings of speech, music, or ambient noise.

Authentication
The process of confirming the origin and integrity of digital content.

Biometric data
Unique physical or behavioural characteristics used to identify individuals (e.g. facial recognition, voice patterns, etc.).

Chain of custody
Documentation of the handling of evidence to ensure integrity in legal and investigative contexts.

CNN (Convolutional neural network)
A deep learning model commonly used in image and video recognition tasks, that is especially effective for analysing visual data.

Content moderation
Processes and systems used to monitor, evaluate, and act on user-generated content online.

Cybercrime
Criminal activities carried out using computers or digital networks.

Deepfake
Media – audio, images, or videos – generated or heavily altered using artificial intelligence to convincingly depict events or people that never occurred.

Detection tools
Software or methods used to identify manipulated or synthetic media.

Digital forensics
The practice of recovering and investigating material found in digital devices.

Disinformation
False information deliberately shared to mislead or manipulate.

Forensic analysis

Systematic and scientific examination of digital or physical evidence to identify, verify, or explain its origin and integrity, often for use in an investigation or court.

Forensic
Related to the use of scientific or technical analysis in the investigation and examination of evidence.

GAN (Generative adversarial network)
A type of AI model that generates realistic synthetic data, such as images, videos, or audio, by training two neural networks against each other.

Image
A single visual representation in digital form, such as a photograph or graphic.

Machine learning
A subset of AI involving algorithms that improve performance through experience or training.

Misinformation
False or misleading information shared without intent to deceive.

Multimedia forensics
The examination of digital audio, videos, and images to detect manipulation, verify authenticity, and ensure reliability as evidence.

Natural language processing (NLP)
A branch of AI focused on enabling computers to understand, interpret, and generate human language.

Neural network
A type of machine learning model inspired by the human brain made up of layers of interconnected nodes ("neurons") that process data and learn patterns from them. Neural networks are the foundation of many AI systems, including those used for image, audio, and language analysis.

Provenance
The origin and history of a digital file, including how and when it was created, modified, or transmitted. Often associated with a chain of evidence where the handling of the article has been recorded.

RNN (Recurrent neural network)
A neural network model designed for processing sequential data, such as audio or text, by retaining memory of previous inputs.

Shallowfake
Media that have been manually edited using basic tools (e.g. cutting, splicing, speed changes) to change context or meaning, without AI involvement.

Verification
Processes used to confirm the authenticity of digital content.

Video

Digital recordings combining moving images and often sound used to visually represent actions, events, or scenes.

INTERPOL

ABOUT INTERPOL

INTERPOL's role is to enable police in our 196 member countries to work together to fight transnational crime and make the world a safer place. We maintain global databases containing police information on criminals and crime, and we provide operational and forensic support, analysis services, and training. These policing capabilities are delivered worldwide and support four global programmes: financial crime and corruption; counter-terrorism; cybercrime; and organized and emerging crime.

OUR VISION: "TOGETHER AGAINST CRIME"

In a world where crime knows no borders, it is clear that collective action is essential to combating crime effectively. INTERPOL serves as a unifying force that brings together police and stakeholders from around the globe with a shared aim: to tackle crime and create a safer world.