INTERPOL

BEYOND ILLUSIONS

UNMASKING THE THREAT OF SYNTHETIC MEDIA FOR LAW ENFORCEMENT

JUNE 2024

BEYOND ILLUSIONS

UNMASKING THE THREAT OF SYNTHETIC MEDIA FOR LAW ENFORCEMENT

Table of Contents

Foreword

In the ever-advancing landscape of technology, particularly with the rise of Artificial Intelligence (AI), synthetically generated media has emerged as a very impactful development for law enforcement. Although not a new technology, the recent rise of generative AI platforms and deepfake-as-a-service offerings has led to a surge in synthetic media variations, now easily produced and accessible online worldwide.

While synthetically generated media offers numerous positive applications across society, it also presents significant challenges. The accessibility and affordability of various AI platforms have enabled criminals to exploit this technology, leveraging synthetic media to propagate and perpetuate criminal activities. This poses considerable challenges for the global law enforcement community. Recognizing the diverse forms of synthetic media and its associated challenges, INTERPOL is committed to exploring this dynamic landscape to support our member countries in addressing both current and future threats posed by synthetic media.

Developing robust forensic and investigative capabilities is crucial for detecting and differentiating between genuine and manipulated media. This requires a multi-stakeholder approach involving member countries, industry stakeholders, and academic institutions to pool our knowledge, and share expertise and advancing technologies. Only through our collective efforts can we mitigate the adverse effects of synthetic media and ensure the sustained integrity and reliability of global law enforcement operations.

To support this endeavor, INTERPOL has developed this background paper to offer a comprehensive analysis of this technology, delving into the potential opportunities offered by it and the challenges it presents for law enforcement.

I trust that this background paper will serve as a cornerstone for INTERPOL in further exploring the synthetic media landscape and assisting member countries in advancing and innovating policing.

Jürgen Stock

INTERPOL Secretary General

Acknowledgement

The successful completion of this effort was made possible through extensive collaboration. There were many collaborators involved in the creation of this background paper on synthetic media and its implication for law enforcement. First and foremost, INTERPOL would like to thank the participants of the INTERPOL Metaverse Expert Group that inspired the formation of the synthetic media expert group to discuss current and upcoming threats and support the creation of this document.

INTERPOL would like to extend a special thanks to Mark Evenblij, Brandon Epstein, Paul Warren-Tape, Matthew Adams, Martino Jerian and Manon den Dunnen for actively contributing to the structure and content of this document and helped shape this paper. I would like to also thank Ananya Das, Christopher Church, Fabio Bruno, Julie Tomaszewski, Janani Nair, Priscilla Cabuyao, Toshinobu Yasuhira, Wookyung Jung, Lindeberg Leite, Mike Price, Parya Lotfi, Mark Nutall, Scott Landman, Jan Collie, Giorgio Patrini and Julia Absalyamova who peer reviewed this document, which greatly assisted in adding invaluable contributions and additional insights for this background paper and were instrumental in filling in the knowledge gaps I extend sincere gratitude to all the experts for their invaluable insights and contributions.

INTERPOL stands ready to continuously explore the opportunities and challenges that synthetic media presents to law enforcement and support member countries in understanding, investigating and applying digital forensics to this emerging threat.

Madan Oberoi
INTERPOL Executive Director of Technology and Innovation

Executive Summary

With the rapid advancement of AI technology, synthetic media is becoming an influential mode of content. Its ease of availability has enabled its exploitation by criminals for malicious activities. Currently, synthetic media files possess the ability to deceive human perception, making it challenging to determine their authenticity. As AI continues to evolve, these artificially generated media will become increasingly sophisticated, posing significant challenges for law enforcement agencies in conducting forensic analyses of such content.

This background paper on synthetic media and its implications for law enforcement, offers a thorough introductory understanding to this emerging technology. Based on the research and analysis of the current synthetic media landscape, the main findings of this paper that are needed to be put forward for law enforcement officers include:

1. Synthetic media is a new and emerging field and tackling the threats posed by it requires dynamic research and analysis of the landscape to understand the technology and find detection solutions.

2. To combat this threat effectively, it is crucial for law enforcement agencies to gain a comprehensive understanding of the multifaceted nature of synthetic media, including its creation, distribution, and potential impact.

3. Law enforcement requires a holistic solution to investigate digital media files efficiently and effectively in this synthetic content era.

4. The rise and increasing ease of AI generated content calls for a collaborative approach between global law enforcement, private sector, and academia to stay at the forefront of synthetic media forensics and investigations through ongoing vigilance, continuous technological advancements, and international collaboration.

In order to assist law enforcement in member countries and enhance collaboration between law enforcement, industry and academia, INTERPOL has drafted this background paper to explore and understand the synthetic media landscape and tackle the threats posed by it. Based on the findings, INTERPOL will continue to work closely with various stakeholders in this endeavor.

1. Introduction

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.

Synthetically generated media has a multitude of positive applications across society, law enforcement, and the private sector. It serves as valuable training material, enables the anonymization of individuals, and consistently delivers high-quality content. While synthetic media content is not a new technology, the recent proliferation of user-friendly generative AI platforms and deepfake-as-a-service offerings has resulted in a surge of various synthetic media variations and use. These can now be produced with minimal effort and are accessible to a broad internet user base worldwide. This sudden rise of synthetic content on the internet has emerged as a significant concern requiring global law enforcement attention.

Addressing the threats of synthetic media requires a multi-faceted approach. For instance, it may be difficult to detect and assess the veracity and authenticity of legal documents or frameworks drafted by AI platforms. Law enforcement agencies need to adapt their investigative approaches to detect and verify the authenticity of media content, as well as collaborate with experts in the field of AI and digital forensics to combat the misuse of synthetic media effectively.

2. Types of Synthetic Media and the Technology Behind

Various categories of synthetic media have emerged as a transformative force in the digital realm, altering how we generate and engage with content. These diverse forms of artificially created or modified media span a broad spectrum of innovative uses, each with its own unique potential and impact. Whether it is in the domain of images, sound, videos, or text, these creations rely on advanced technologies for their generation. With the capacity to imitate and, in some cases, surpass human-produced content, synthetic media represents a dynamic frontier with a potential for both positive and concerning implications. Experts believe that 90% of the content available online will be synthetically sourced by 2025[1]. Some of the key types of synthetic media discussed in this paper are outlined below.

2.1 Deepfakes

Representing a specific type of synthetic media, deepfakes are created using sophisticated AI algorithms to depict individuals in highly realistic, and often deceptive, situations that never occurred. These advanced AI models are capable of creating highly convincing visual and audio content, often featuring individuals in fabricated scenarios or expressing statements they have never stated.[2]

[1] Giardina, C. (2023, January 8). CES: Could 90 Percent of Content Be AI-Driven by 2025? The Hollywood Reporter. https://www.hollywoodreporter.com/movies/movie-news/ces-ai-sag-aftra-1235290431/

[2] Goodfellow, I. J., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (https://arxiv.org/pdf/1406.2661.pdf)

The emergence of deepfakes has generated profound implications across various domains, raising both interest and concern in equal measure. With reference to its usage, deepfakes have unlocked new horizons in the realm of special effects, enriching the entertainment industry with lifelike CGI[3] and offering unparalleled creative potential. Conversely, the nefarious use of deepfakes for misinformation, identity theft, and manipulation has raised significant ethical and security concerns, underscoring the urgent need for safeguards and countermeasures.

The principal algorithms used in the generation of deepfakes are:

Algorithm	Description and Concept	Application in Deepfakes
Generative Adversarial Networks (GANs)	GANs are composed of a Generator and a Discriminator[4]. GANs engage in a creative duel where the generator generates new media based on the training data and the discriminator acts a gatekeeper, making sure the generated media looks real or fake.	GANs are mainly used to generate digital clones of a certain individual by mimicking facial expressions in videos. (Figure1 (1))
Diffusion Models	Diffusion models provide a guided approach to generating diverse and high-quality deepfakes by gradually transitioning between distributions of input data points. This process preserves sample diversity and facilitates the computation of conditional probabilities, contributing to the creation of realistic deepfakes.[5]	Diffusion Models are used to create Picture to Picture transformations that allow for smooth transition between image distributions.(Figure1 (2))
Autoencoders	Autoencoders are a type of neural network that are trained to efficiently encode input data, capturing essential features while discarding irrelevant information.[6] This encoded representation is then used to reconstruct the original input, enabling autoencoders to generate deepfakes by seamlessly replacing facial features in videos while retaining image quality and appear authentic.	Autoencoders excel at face swaps using its ability of retaining image quality while seamlessly replacing facial features in videos deepfakes. (Figure1 (3))

Table 1: Technology used for the generation of Deepfakes

[3] CGI – Computer Generated Imagery. Computer Generated Imagery are a type of visual effects in which scenes, effects and images are synthetically generated using a computer software.

[4] Luca Guarnera et al. (2023) Level Up the Deepfake Detection: A Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models (https://arxiv.org/pdf/2303.00608.pdf)

[5] Aditya Ramesh et al. (2022) Hierarchical Text-Conditional Image Generation with CLIP Latents (https://arxiv.org/pdf/2204.06125.pdf)

[6] https://www.ibm.com/topics/autoencoder

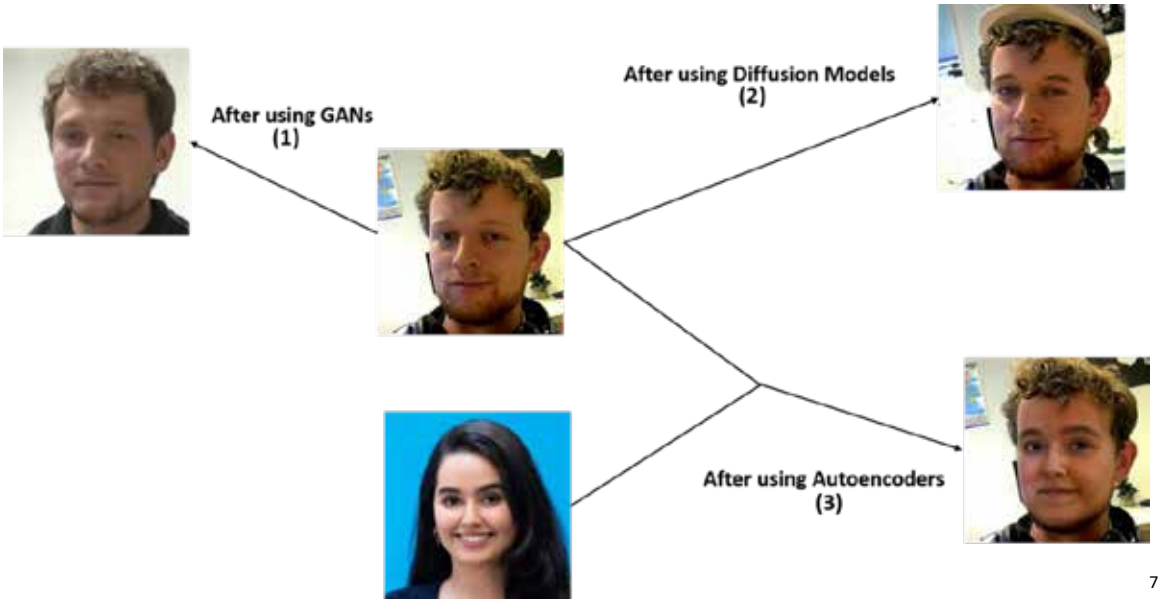

Figure 1: The Different Deepfakes produced with different models.[8]

2.2 Synthetic Audio

Voice synthesis is a distinct niche within the synthetic media landscape that has emerged with the rise of generative AI. It involves the creation of lifelike speech patterns and audio content, often mimicking the vocal characteristics of specific individuals, or generating a totally new tone or style of speech. In its constructive application, voice synthesis has revolutionized the accessibility of technology, aiding individuals with speech impairments, and enhancing the capabilities of virtual assistants with more human-like interactions.

On the other hand, voice synthesis could be misused for manipulation of audio recordings, impersonations, and other deceptive practices. These nefarious activities forecast the critical need for security measures and ethical guidelines to safeguard against misuse. The ability to generate convincing synthetic voices has deepened concerns about identity theft, fraud, and the credibility of audio evidence, necessitating vigilance and countermeasures to uphold the integrity of auditory content in a digitally evolving world.

[8] Image Contributed by DuckDuckGoose (AI Detection Services | Verifying ID & Combatting Fraud (duckduckgoose.ai))

Some of the main AI algorithms utilized for creating synthetic audio are:

Algorithm	Description and Concept	Application in Synthetic Audio
Recurrent Neural Networks (RNNs)	RNNs detect patterns in sequential data to synthesize audio files.[9] By training on voice recordings, RNNs learn the nuances of human speech, capturing phonetic, prosodic, and expressive elements to produce accurate and personalized synthetic voices.	RNNs are used for voice synthesis, RNNs generate synthetic voices with human-like quality, aiding in applications such as virtual assistants and personalized speech synthesis for individuals with speech impairments.
WaveGAN and Variational Autoencoders (VAEs)	WaveGAN utilizes principles from GANs to generate raw audio slices, producing remarkably realistic audio samples.[10] VAEs offer potential for speech generation, musical composition, and sound design, bridging the gap between computational efficiency and high-quality audio synthesis.[11]	WaveGAN transforms image-processing models into audio synthesis tools, generating diverse and high-quality audio content. VAEs enable unconditional generation of speech and music, pushing the boundaries of audio synthesis with enhanced computational efficiency and exceptional quality.

Table 2: Algorithms used to synthesize audio

2.3 Generated Text

Text generation represents a distinct area of synthetic media, enabling the automated creation of written content with remarkable authenticity. It is the task of generating text with the goal of appearing indistinguishable to human-written text, more formally known as Natural Language Generation (NLG). Text generation plays a pivotal role in the dynamic landscape of synthetic media, contributing to the creation of diverse and engaging content. The current most advanced methods for text generation, including platforms like ChatGPT or BART AI, enable the generation of human-like and contextually relevant text across various domains. In the realm of synthetic media, these models empower e.g. content creators to produce realistic and coherent narratives for virtual characters, dialogues for interactive storytelling, and even generate new article or blog posts with minimal human intervention.

[9] Robin M. Schmidt (2019). Recurrent Neural Networks (RNNs): A gentle Introduction and Overview (1912.05911.pdf (arxiv.org))
[10] Chris Donahue et al. (2018) Adversarial Audio GANs (https://arxiv.org/pdf/1802.04208.pdf)
[11] Antoine Caillon et al. (2021) RAVE: A variational autoencoder for fast and high-quality neural audio synthesis (https://arxiv.org/pdf/2111.05011.pdf)

The principal algorithms used in the generation of text are:

Algorithm	Description and Concept	Application in Generated Text
Recurrent Neural Network (RNNs)	RNNs generate text with their ability to process and retain past information[12], overcoming challenges like the vanishing gradient problem[13]. Modern RNNs excel in character-level language modelling, offering potential for revolutionizing text compression and enhancing accessibility for individuals with physical disabilities.	RNNs are used for generating text with rich vocabulary, intricate grammatical structures, and diverse proper names. Handles tasks like balancing parentheses and quotes across long distances, demonstrating proficiency in understanding and producing complex language structures. Multiplicative RNNs (mRNNs)[14] provide an edge in processing and generating intricate text.
Transformer Models	Transformer models within the context of data-to-text generation, transform structured data into readable narratives. Traditionally, this involved two steps: deciding what to say and how to say it, but with sequence-to-sequence (S2S) learning, models now perform both tasks simultaneously. Challenges include working with structured data and summaries to generate responses, which are more complex than other text generation tasks.[15]	Transformer models are utilized for creating summaries based on the underlying structure of input data. Performs data-to-text generation tasks efficiently, addressing challenges in producing responses from structured data and summaries.
Generative Pre-Trained Transformer (GPT)	GPT models is very good for generating human-like and contextually coherent text by predicting the next word based on preceding context. Trained on extensive text and datasets, they demonstrate proficiency in various natural language processing (NLP) tasks, including text completion, question answering, and language translation.[16]	GPT model is used for generating text that sounds natural and contextually relevant, owing to their understanding of relationships and patterns within language. Versatile enough for fine-tuning to specific tasks or domains, enhancing performance in generating text tailored to particular contexts.

Table 3: Algorithms used for text generation

[12] Ilya Sutskever et al. (2011) Generating Text with Recurrent Neural Networks (https://www.cs.toronto.edu/~jmartens/docs/RNN_Language.pdf)
[13] Vanishing Gradient Problem - Vanishing gradient problem is a phenomenon that occurs during the training of deep neural networks, where the gradients that are used to update the network become extremely small or "vanish" as they are backpropagated from the output layers to the earlier layers. (engati.com)
[14] mRNN - A Multiplicative RNN (mRNN) is a type of recurrent neural network with multiplicative connections. (paperwithcode.com)
[15] Li Gong et al. (2019) Enhanced Transformer Model for Data-to-Text Generation (https://aclanthology.org/D19-5615.pdf)
[16] Open AI (2023) GPT-4 Technical Report (https://cdn.openai.com/papers/gpt-4.pdf)

2.4 Synthetic IDs

Synthetic IDs are artificial identity documents to produce images resembling genuine and authentic identification documents. These IDs are designed to replicate the appearance of legitimate identification papers, including driver's licenses, passports, or other official forms of identification.

The primary aim of generating synthetic IDs is to support the volume of data available for training deep neural networks, specifically for the purpose of detecting and preventing fraudulent identity documents. This approach assists in enhancing fraud detection mechanisms while simultaneously addressing concerns related to privacy and the sensitive nature of personal identity documents.[17]

Synthetic IDs can be generated using the following principal algorithms:

Algorithm	Description and Concept	Application in Synthetic IDs
Computer Vision and GANs	Computer vision algorithms and GANs work together to generate synthetic ID card images for enhancing fraud detection systems. GAN models simulate genuine ID card examples and introduce texture noise, while computer vision techniques enable automatic splicing to create composite scenarios, thereby expanding datasets for training fraud-detection networks.	Computer Vision and GANs are employed to produce diverse classes of ID cards, simulate genuine examples, introduce texture noise, and create composite scenarios, thereby expanding datasets for fraud-detection network training and improving the accuracy of identifying fraudulent activities.
Diffusion Probabilistic Models	Diffusion Probabilistic Models, especially Large diffusion models are employed as generative models, to generate and augment document data, address challenges in accurately representing complex identity documents. A customized pipeline combines optical character recognition, LayoutLM[18] understanding, face detection, and text layout generation to create fake identity documents with improved fidelity.	Diffusion Probabilistic Models are utilized to generate and augment document data, particularly identity documents, by employing a customized pipeline that integrates optical character recognition, layout understanding, face detection, and text layout generation, thereby improving the fidelity of synthetic identity documents.

Table 4: Algorithms for Synthetic ID Generation

[17] Daniel Benalcazar et al. (2022) Synthetic ID Card Image Generation for Improving Presentation Attack Detection (https://arxiv.org/pdf/2211.00098.pdf)
[18] Yiheng Xu et. al. (2019). LayoutLM: Pre-training of Text and Layout for Document Image Understanding (https://arxiv.org/abs/1912.13318)

(A) *(B)*

Figure 2: Synthetic IDs: (A) Driver's License (B) A passport[19]. These synthetic IDs look very close to the authentic IDs and often can be mistaken as original or even pass identity verification tests

3. Enabling and Enhancing Law Enforcement Capabilities

Synthetic media offers a range of benefits for law enforcement, revolutionizing various aspects of their operations and strategies.

3.1 Montage Generation

Synthetic media may also be used to translate eyewitness descriptions into visual representations of potential suspects: through analysis and interpretation of provided details, such as facial features or distinct physical traits. This aids law enforcement investigations by providing leads and helping to identify and locate potential suspects involved in criminal activities. By generating accurate and detailed suspect images, synthetic media significantly supports law enforcement efforts, streamlining investigations for more effective outcomes.[20]

3.2 Undercover Operations and Covert Surveillance

Synthetic media significantly helps law enforcement capabilities through its roles in undercover operations and covert surveillance. Primarily, in undercover operations, this technology serves as a cornerstone for crafting and adapting personas of undercover officers. By leveraging synthetic media, law enforcement agents can swiftly modify appearances, incorporating fictional injuries, scars, tattoos, or other alterations necessary to sustain cover identities. This adaptability ensures the maintenance of plausible and dynamic personas critical for successful undercover assignments.

Moreover, in the realm of covert surveillance, synthetic media plays a key role in supporting law enforcement by assisting in the creation of realistic backstories for undercover officers. This technology contributes to the development of believable and authentic personas for undercover

[19] Image Contributed by IDVerse. (Home - IDVerse)
[20] https://www.researchgate.net/publication/342978037_Suspect_Face_Generation

operatives, enhancing the officers' capabilities to operate incognito within diverse environments. By providing the means to establish credible backstories, synthetic media augments the effectiveness and adaptability of covert surveillance strategies employed by law enforcement agencies.[21]

3.3 Training and Public Engagement

Synthetic media plays a significant role in augmenting law enforcement capabilities, particularly in training, public engagement, and officer anonymity. Within the domain of training, law enforcement agencies leverage synthetic media to create highly realistic and immersive scenarios. By simulating authentic training environments, officers can enhance their preparedness to handle diverse and complex situations they might encounter in the field. This enables them to refine their skills, decision-making, and response strategies in a controlled yet realistic setting.

Furthermore, synthetic media finds utility in event reconstructions, aiding law enforcement in recreating crime scenes or incidents for investigative purposes or public appeals. By recreating scenarios using synthetic media, law enforcement can potentially gather valuable evidence, seek public assistance, and increase community engagement in solving cases.

4. Potential Challenges for Law Enforcement

The threat that synthetic media poses to law enforcement is primarily related to its potential for misuse and deception. Here are a few threats it can pose:

4.1 Evidence Authentication

The rise of synthetic media brings forth significant challenges for evidence authentication in investigations. Determining the authenticity of media content becomes a complex endeavor due to the sophisticated manipulation capabilities. Traditional methods of visual and audio analysis may no longer be sufficient to identify manipulated content, thereby requiring specialized expertise and the utilization of advanced technologies for forensic analysis. This places additional burdens on investigative bodies and legal authorities to keep pace with the evolving landscape of synthetic media and its potential implications for the justice system.

Challenge to Legal Proceedings

Synthetic media can be used to create fake evidence, such as forged videos or manipulated images, which can be presented as genuine in legal proceedings. This could lead to wrongful accusations, false alibis, or the undermining of an investigation. In an era where digital evidence carries substantial weight, the potential for malicious actors to exploit synthetic media for their advantage is a growing concern. Instances of fake alibis, fabricated crime scenes, or manipulated testimonies can emerge, casting doubt on the integrity of the legal system and the pursuit of justice. Consequently, the rise of

[21] Kelly W Sundberg and Christina M Witt (2019); Undercover operations: Evolution and modern challenges. Undercover operations: Evolution and modern challenges | Journal of the Australian Institute of Professional Intelligence Officers (informit.org)

synthetic media requires a closer examination of how legal systems adapt to address these emerging threats, ensuring that the pursuit of truth and fairness remains paramount.

The proliferation of synthetic media poses a significant challenge when it comes to proving the authenticity of evidence in legal proceedings. In an environment where fabricated media can be virtually indistinguishable from genuine content, the burden of verifying the veracity of audio, video, or images becomes increasingly complex. This challenge is compounded by the emergence of the "Liars Dividend," a phenomenon where defendants claim that genuinely authentic media has been artificially generated, creating a climate of doubt and uncertainty. As a result, judicial systems face the pressing need to adapt to this evolving landscape, developing robust methods and standards for evidence authentication to uphold the integrity of legal proceedings.

4.2 Victim Identification

The emergence of synthetic media has given rise to a distressing concern regarding non-consensual explicit content. Malicious actors can exploit this technology to create compromising content featuring individuals without their knowledge or consent. In such scenarios, identifying and providing support to the victims becomes notably intricate, as the line between reality and synthetic fabrication blurs. Addressing these issues requires heightened vigilance, legal frameworks, and support systems for potential victims to ensure their well-being and protection in an increasingly digital and interconnected world.

Impersonation

Synthetic media enables highly convincing impersonations of individuals, presenting a concerning challenge. For instance, an individual's voice can be synthesized with remarkable accuracy, making it exceptionally challenging to distinguish between a genuine and manipulated phone call or audio recording. This capability can be exploited for a range of malicious purposes, from carrying out extortion schemes to disseminating false information, thereby amplifying concerns about identity verification and trust in various aspects of communication and digital interactions. In an age where authenticity is increasingly fragile, addressing the potential for impersonation is a crucial aspect of managing the risks associated with synthetic media.

- **AI Voice Scams**

AI voice scams ingeniously leverage technology to replicate the voice of a victim's family member, friend, or even impersonate customer service representatives, all with the intent to deceive. Perpetrators utilize these fabricated voices to manipulate victims into revealing sensitive personal information or coercing them into sending money. The scam operates by exploiting trust and familiarity, leading unsuspecting individuals to believe they are interacting with someone they know or a legitimate service provider, amplifying the success of the ruse.

- **Bypassing liveness checks**

A liveness check tests whether the person in front of the camera is the actual user or if the user is attempting to spoof an identity that is not theirs. Active liveness is the most common and consists in asking the user to perform certain actions like blinking, smiling, or following movement instructions with their head. Despite its widespread adoption, active liveness checks are weak against attacks by

Deepfakes. The reason is that real-time Deepfakes can reproduce faithfully facial landmark movements of the attackers. Hence, even if facial traits e.g. eye shape and color, the height of the cheekbones, the shape of mouth do change, the attacker can move their head left and right and smile, easily following the active liveness instructions.

Know Your Customer (KYC) is the practice of verifying an individual's identity in compliance with laws and regulations, primarily for anti-money laundering purposes. In response to the pronounced inability to detect and prevent online fraud, the KYC process has been introduced to reduce instances of illegal transactions. Due to its sensitivity for businesses, KYC is regulated by respective national and international government agencies. KYC procedures are necessary when opening an account with a bank or when conducting financial transactions of any sort. The possibility of bypassing KYCs with aid of synthetic media may pose a huge threat to the online liveness check procedures, as it enables chances of impersonation and thus compromises the reliability of the identity verification process.

In a recent incident, it was reported that a multinational corporation fell victim to a sophisticated deepfake scam, resulting in a huge financial loss of $25.6 million. With the use of advanced synthetic media technology, perpetrators digitally replicated high-ranking company officials, including the Chief Financial Officer, within a simulated video conference setting. The deepfake representations convincingly mimicked the appearance, voice, and mannerisms of genuine personnel, fostering an illusion of authenticity. Under the guise of legitimacy, the victim was deceived into executing a series of 15 transfers totaling millions of dollars to multiple bank accounts. Despite initial skepticism, the victim's compliance was facilitated by the seamless interaction and seemingly genuine appearance of the deepfake personas.[22]

4.3 Misinformation and Propaganda

Synthetic media's capacity to generate convincing news articles, videos, or social media posts raises substantial concerns related to the dissemination of false information and the manipulation of public opinion. This creates formidable challenges for law enforcement agencies, which heavily depend on accurate and reliable information to conduct effective investigations and ensure public safety. The proliferation of synthetic media exacerbates the complexity of discerning fact from fiction, underscoring the pressing need for robust mechanisms to verify the authenticity of digital content and safeguard against the potential consequences of misinformation and propaganda in an information-driven society.

Simultaneously, the rise in use of text generation tools also raises serious concerns. These ranges from questions of plagiarism and intellectual property rights, to ethical concerns, as it becomes crucial to ensure responsible use to prevent misinformation, deepfakes, or manipulation of public opinion. Large Language Models (LLM) are vastly used to generate content across social media. Unsupervised use of this content may lead to the spread of misinformation and propaganda. Social media is a breeding ground for such false information, attracting such malicious content to reach a larger audience.[23]

[22] https://economictimes.indiatimes.com/industry/tech/hong-kong-mnc-suffers-25-6-million-loss-in-deepfake-scam/articleshow/107465111.cms?from=mdr

[23] https://www.cpahq.org/media/ivih25ue/handbook-on-disinformation-ai-and-synthetic-media.pdf

4.4 Privacy Concerns

Synthetic media's capability to create explicit or compromising content, featuring individuals who may have never engaged in such behavior, escalates profound privacy concerns. This technological potential can be exploited for malicious purposes, including harassment, blackmail, or causing significant harm to an individual's reputation. The ease with which synthetic content can be produced and disseminated amplifies the need for robust privacy protection measures and legal safeguards to mitigate the potential harm that could be inflicted upon individuals.

For instance, in circumstances involving deepfake-fueled extortion, perpetrators engage victims in interactions until they obtain explicit images or videos. The extortion scheme relies on these manipulated, synthetic media as the leverage for blackmail, bypassing the need for establishing rapport or befriending the victim. Once in possession of these compromising deepfake images or videos, the culprits coerce victims into complying with their demands, threatening to expose or distribute the falsified content if their demands are not met.

These synthetic IDs by impersonation of legitimate personal data mixed with synthetic information, allow criminals to create false identities that could be exploited by fraudsters to secure credit lines, apply for government benefits, intercept tax returns, and participate in various financial activities, resulting in significant monetary losses for financial institutions and government agencies.24

4.5 Forensics in the Synthetic Media Era

A key element of the challenge facing forensic experts lies in the need to understand the AI media generation process and integrate this advanced knowledge into the forensic identification steps. While automated tools offer valuable assistance, they are not entirely adequate. To conduct effective forensic analysis amid the ever-evolving landscape of AI-generated media, it is essential to combine the use of these tools with rigorous analysis provided by experienced forensic professionals. This balanced approach ensures an unbiased and interpretable examination.

[24] https://legal.thomsonreuters.com/en/insights/articles/synthetic-identity-fraud

5. Investigative Techniques and Forensic Analysis

5.1 Collection of Evidence

Investigative techniques and forensic analysis are vital in combating the growing unlawful synthetic media content. For a seized evidence file, the following investigative techniques can be applied to check the authenticity of an evidence file.

Source Verification

This technique involves verifying the origin and authenticity of a particular piece of synthetic media. Investigators assess the credibility of the source, the techniques used to create the content, and any potential motives for dissemination. This multifaceted approach requires technical expertise, adaptability to evolving methods, and advanced technologies to detect manipulated content and mitigate harm.

Metadata Analysis

Metadata analysis by supplying vital contextual information aids in the realm of synthetic media investigations. This includes details about the source, creation date, and editing history, providing investigators with the means to trace the origins and authenticity of potentially deceptive or harmful content. However, it is important to note that in certain scenarios, particularly when the media content has been transmitted over various applications, the metadata may be compromised or incomplete. This inherent limitation highlights the need for a multi-faceted approach, combining metadata analysis with other investigative techniques to ensure a comprehensive examination of synthetic media.

Reverse Image Searching

Reverse image searching is a process of taking an image and searching for similar or exact matches on the internet. It plays an important role in the arsenal of investigative tools when determining the authenticity or tracing the origins of media content, especially when dealing with morphed or manipulated features. When the original, unaltered version of the content can be found online, reverse image searching can quickly unveil the deception, assisting in the identification of manipulated or synthetic media.

Linguistic Analysis

Delving into the linguistic aspects and writing style embedded within synthetic media is a strategic approach for unraveling its true origins. The unique language patterns, cultural allusions, and even grammatical anomalies can serve as valuable clues in discerning the authenticity of the content. Experts employ linguistic analysis to ascertain whether the media in question is a product of manipulation or an authentic piece. This examination of language serves as a linguistic fingerprint that can help investigators determine whether the content aligns with the expected linguistic traits of the purported source, thereby aiding in the authentication of multimedia materials.

5.2 Emerging Technology Solutions for Synthetic Media Detection

Synthetic media refers to any kind of media content that has been artificially created or manipulated using advanced technologies. Some of the detection solutions currently used are illustrated below.

Deep Learning Models

Deep learning models in detecting synthetic IDs and synthetic media leverage various techniques and algorithms and follow the concept of “AI is needed to detect AI”. One significant approach involves training models to differentiate between real and AI-generated images or videos by identifying inconsistencies or artifacts left by generative models. These include abnormal textures, synthetic movements, or subtle flaws often overlooked by the human eye but detectable by a well-trained AI system. Moreover, unsupervised learning techniques, like anomaly detection algorithms, prove useful in flagging potential forgeries as generative AI fraud introduces atypical patterns into the data.

In addition to identifying anomalies, digital forensic techniques are applied to analyze the metadata of images and videos, examining format, compression, and noise distribution. Generative AI typically leaves distinct digital fingerprints that forensic tools can detect. To stay ahead, incorporating adversarial training exposes detection models to the latest AI-generated fraud, enabling them to learn and adapt to new and evolving generative techniques.

Further combining multiple AI models to analyze diverse aspects of the data, such as visual and temporal features enhance detection rates. For instance, a deepfake video might display realistic visuals but may lack synchronization between the facial and verbal expressions, detectable by an AI trained to recognize such discrepancies. Moreover, AI systems cross-reference content with known databases of authentic images, videos, or documents, revealing discrepancies that could signify fraudulent activity. This proves particularly effective in document verification, where details are cross-checked against official records, enhancing the precision of synthetic media and synthetic ID detection methods.

Explainable AI

The differentiation between authentic and synthetic media is becoming critically important. Deepfake detection leveraging AI presents a promising solution to manage the increasing volume of synthetic content. However, the inherent "black box[25]" nature of these AI systems hampers the transparency of their decision-making processes, raising the question, "Why is this image classified as fake?" As the average individual's ability to distinguish between sophisticated synthetic images and reality diminishes, the necessity for Explainable AI (XAI) in this domain grows. Some of the explainable deepfake detection approaches include the following.

[25] Black Box - A black box refers to a system whose behavior has to be observed entirely by inputs and outputs. Even if the internal structure of the application under examination can be understood, the tester chooses to ignore it. Black box testing assesses a system solely from the outside, without the operator or tester knowing what is happening within the system to generate responses to test actions. (techtarget.com)

Method	Description
Class Activation Mapping (CAM)	CAM offers a generalized approach to enhance interpretability of AI decisions. Directing focus to specific areas within deepfakes, CAM offers a valuable entry point analysis and understanding to the AI reasoning.
Deepfake Detection Typology	Deepfake detection typology helps identify the deepfake models which can provide insights into origins or creator identity, akin to tracing cybercriminal trends.
Digital Fingerprinting	The digital fingerprinting technique traces deepfakes to their originating GAN through unique digital fingerprints. While such fingerprints can be masked or removed, their presence unequivocally indicates content manipulation.
Biometric Analysis	Biometric analysis with the utilization of physiological signals like heartbeat patterns and eye reflections aid in affirming human authenticity. Despite potential for spoofing, when applied correctly, this method can be pivotal in flagging deepfakes—such as in cases where high-resolution images lack expected biometric data.

File Structure Analysis

File Structure Analysis is a non-content-based authentication approach that leverages the way a file is constructed at a binary level. Because it does not rely on displayed imagery, the quality of the video does not affect the analysis; meaning high quality or low-quality synthetic media is equally detectable. File Structure Analysis also does not rely on metadata values which can easily change due to intentional manipulation or when transmitted between devices. As a deterministic, transparent and repeatable approach, the results of File Structure Analysis are explainable as to how they are generated and can be evaluated by the end user. File Structure Analysis can also be used to prove that a file is camera original and attributed to a specific brand and model of device, allowing for not only the detection of synthetic media but the authentication of "real" imagery. While File Structure Analysis will accurately and reliably identify camera original versus synthetic files at scale in an automated process, it will not identify the specific content that could have changed.

Biological Signals

Biological signals, such as photoplethysmography (PPG) cells[26], can assist in discerning between authentic and synthetic media. They are used in a smart system that spots different kinds of generated models used in synthetic videos. These signals have patterns that help show where videos might have been altered by highlighting out the changes from the real biological data. These signals are harnessed to power a classification network, enabling the detection of generative models employed in videos. The intricate spatiotemporal patterns within these biological signals serve as a crucial representation of residuals. These patterns effectively unveil manipulation artifacts by segregating them from the authentic biological signals. Given that biological signals remain absent in deepfakes, their absence yields distinct signatures discernible in the generative noise. Consequently, these biological signals serve as an indispensable projection, representing the residuals within a known dimension. This

[26] Photoplethysmography (PPG) cells - Photoplethysmography (PPG) measures the amount of light absorbed or reflected by human tissues. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8950880/)

projection allows for an exploration of unique signatures per model, enabling the identification and differentiation of synthetic media from authentic content.[27]

Statistical Analysis at a Pixel Level

Statistical analysis at a pixel level involves scrutinizing decoded video pixels or coefficients to identify irregularities that might indicate the presence of synthetic media. This process focuses into the details of video data, examining statistical patterns and anomalies that could be attributed to the generation of synthetic content. For instance, in this method, the artifacts generated by convolutional neural networks is observed during their processing and often visible in the Fourier transform[28] of pixels.

By analyzing the statistical characteristics of decoded video pixels or their transformations, this technique aims to uncover telltale signs of synthetic generation. Anomalies in statistical patterns, unusual fluctuations, or irregularities in pixel transformations may signify the manipulation or artificial generation of the media. These irregularities can range from subtle variations to distinct patterns that deviate from what would be expected in naturally captured or authentic content.

Geometrical and Behavioral Analysis

Geometrical and behavioral analysis plays a crucial role in detecting synthetic media by careful inspection of various inconsistencies that might be imperceptible to the human eye. These analyses look into elements like shadows, reflections, perspective, proportions, and other geometrical attributes to unveil discrepancies that could indicate synthetic generation.

For instance, when examining shadows and reflections within synthetic media, an in-depth analysis focuses on their consistency and adherence to natural laws. Synthetic content may inadvertently introduce inconsistencies in how shadows fall or how reflections interact with the environment, deviating from the expected behavior seen in authentic visual content. By systematically analyzing these elements, anomalies such as incorrect lighting angles or inconsistent shadow lengths can be identified, indicating potential manipulation.

Similarly, behavioral analysis involves studying the behavior of objects or entities depicted in the media. Synthetic media might exhibit discrepancies in behavioral patterns that depart from expected real-world behaviors. This examination involves assessing how objects move, interact, or behave within the digital environment. Any unnatural or irregular movements, unnatural physics, or peculiar interactions within the synthetic content can signal potential artificial manipulation.

AI-generated Speech Analysis

AI speech detection involves a blend of signal processing techniques and machine learning algorithms to analyze and interpret audio data. Initially, the input speech undergoes preprocessing to extract pertinent features, which can include spectrograms including Mel-frequency cepstral coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs). These extracted features encapsulate both

[27] Umur Aybars Ciftci et al. (2020) How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals (https://arxiv.org/pdf/2008.11363.pdf)

[28] Fourier Transform - The Fourier Transform is an important image processing tool which is used to decompose an image into its sine and cosine components. The output of the transformation represents the image in the Fourier or frequency domain, while the input image is the spatial domain equivalent. (homepages.inf.ac.uk)

frequency and time characteristics of the audio, enabling a detailed representation of speech patterns. These features serve as inputs to deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), adept at learning intricate patterns and variations in speech. By leveraging these diverse feature sets, these models can power speech detection systems.

6. Regulatory and Policy Considerations

Due to the potential implications and risks associated with synthetic media, several regulations and guidelines have been proposed or implemented to address concerns such as misinformation, privacy, copyright infringement, and ethical use.

One example is the deepfake technology, which refers to the use of AI to create realistic synthetic media, often used to manipulate videos or images of individuals. In many countries, laws regarding privacy and defamation can be applied to deepfakes, holding creators accountable for violating privacy rights, spreading false information, or defaming individuals.

6.1 Intellectual Property

The evolution of synthetic media poses significant challenges in copyright ownership. With AI-generated content often blurring the lines of original authorship, determining rightful ownership becomes intricate. Identifying the creator or owner of manipulated content in synthetic media raises questions about copyright claims and protections. The complexities involved in establishing ownership rights in AI-generated works call for a reassessment and potential adaptation of copyright laws to address the novel aspects of content creation in the digital age.

The concept of fair use undergoes scrutiny in the context of synthetic media. Fair use exceptions are subject to debate when assessing the transformative nature of AI-generated content. Differentiating between legitimate transformative works and potential copyright infringement becomes increasingly challenging. The interpretation of fair use laws needs refinement to accommodate the transformative aspects of AI-generated content, ensuring the balance between creative freedom and copyright protection.

The responsible use of AI in law enforcement prioritized the alignment of policing principles, ethical standards, and human rights compliance. With this intent, INTERPOL, in partnership with the United Nations Interregional Crime and Justice Research Institute (UNICRI), published the Toolkit for Responsible AI Innovation in Law Enforcement[29] in June 2023. This toolkit, comprises of seven resources aimed at guiding law enforcement agencies in designing, developing, procuring, and deploying AI technologies in a responsible manner.

Policymakers face the task of adapting existing intellectual property regulations to effectively encompass the nuances presented by synthetic media. Establishing guidelines and standards that govern AI-generated content is imperative to ensure compliance with intellectual property laws. Crafting regulatory frameworks that address the intricate issues of copyright, licensing, trademark

[29] https://www.interpol.int/en/How-we-work/Innovation/Artificial-Intelligence-Toolkit

protection, and fair use within the context of synthetic media creation is essential to safeguard the interests of creators and copyright holders while fostering innovation.

Memo

In many countries, the exploration and utilization of generative AI, such as ChatGPT and Gemini (previously Bard AI), have triggered intricate intellectual property debates under existing legislation. There are countries whose domestic laws permit “information analysis”, allowing the use of copyrighted materials for the training of generative AI models in the training stage.

6.2 Explainable AI

"Explainable AI" is emerging as a crucial factor in AI applications, especially in digital forensics, for detecting synthetic or deepfake media and processing extensive data. To be considered "Explainable AI," it is important for these tools to adhere to the principles outlined by NIST[30] Circular 8312. This entails providing clear, accurate, and understandable explanations for AI-generated outputs, including confidence levels, beyond mere yes/no responses.

Memo

In some countries, within legal contexts, authenticating digital evidence traditionally involves witness testimony. However, AI tools cannot testify, leading to requirements for expert witnesses familiar with AI technology to authenticate evidence in court. Yet, challenges persist in explaining AI processes adequately for authentication purposes. Moreover, some standards assessing the admissibility of novel scientific testimony demand that AI techniques have been tested, published, demonstrated error rates, follow standards, and enjoy wide scientific acceptance. Transparency in AI algorithms becomes pivotal as a lack of explainability may hinder the admissibility of AI-generated results.[31]

[30] NIST - National Institute of Standards and Technology

[31] https://medexforensics.com/2023/10/30/evaluating-the-use-of-ai-in-digital-evidence-and-courtroom-admissibility/

7. Recent Reports on Synthetic Media

With the rise of generative AI, the ease of producing synthetic media has significantly increased which pose a serious concern in terms of their widespread creation and dissemination.

7.1 Deepfakes

- Two men have been founded guilty for using AI to generate exploitative images of children, marking a significant instance in such cases. The court's ruling established a significant precedent by recognizing that sexually abusive content encompasses AI-generated imagery with a "high level" of realism that can resemble real children and minors.[32]

- An investigation has been launched into the use of AI to manipulate images of young girls, altering their clothing and circulating the edited photos.. It was revealed that multiple girls were affected, and some of the preparators have been identified by the police. This stressed the severity of the situation, calling for those responsible to cooperate and expressing fears that the images might be disseminated on websites. This incident highlights the rising problem of digital violence and the pressing need to address it.[33]

- In Quebec, Canada, a man was found guilty for producing synthetic videos of child sexual abuse using deepfake technology. The judge highlighted the detrimental impact of these synthetic images, not only fueling the market for child sexual abuse material but also complicating police investigations and endangering the safety of children online whose identities could be exploited.[34] A Purple Notice[35] was issued to alert INTERPOL member countries on the use of deepfake technology to defraud and extort victims through impersonation scams, online sexual blackmail and investment fraud.

- In 2023, The National Centre for Missing & Exploited Children (NCMEC) received approximately 5000 reports of AI-generated child exploitation media files. NCMEC's CyberTipline 2023 Report[37] highlights the concern of such AI-generated Child Sexual Abuse Material or AIG-CSAM created by bad actors based on a real child or computer-generated children. The report also states that more than 70% of these AIG-CSAM files were from conventional online platforms[38], thus highlighting the inadequate content tracking protocols on these platforms. However, in April 2023, industry leaders like OpenAI, Amazon, Anthropic, Civitai, Google, Meta, Metphysic, Microsoft, Mistral AI, and Stability AI have joined an initiative led by Thorn to adopt 'Safety by Design Principles' to prevent the creation and proliferation of AIG-CSAM[39].

[32] https://edition.cnn.com/2023/09/27/asia/south-korea-child-abuse-ai-sentenced-intl-hnk/index.html

[33] https://edition.cnn.com/2023/09/20/europe/spain-deepfake-images-investigation-scli-intl/index.html

[34] https://www.cbc.ca/news/canada/montreal/ai-child-abuse-images-1.6823808

[35] INTERPOL Purple Notice: To seek or provide information on modus operandi, objects, devices and concealment methods used by criminals.

[36]The National Center for Missing & Exploited Children is a private, non-profit corporation based in the United States

[37] https://www.missingkids.org/cybertiplinedata

[38] https://www.missingkids.org/blog/2024/generative-ai-csam-is-csam

[39] https://openai.com/blog/child-safety-adopting-sbd-principles

7.2 Synthetic Audio

- A woman faced a harrowing ordeal when she received a call from someone who sounded exactly like her grandson, claiming to be in jail and in need of bail money. Believing it to be her grandson, she withdrew the ransom money from the bank and were about to send it when a bank manager intervened. They soon realized the caller had impersonated their grandson, and they had been scammed This incident reflects a growing trend of impersonation scams aided by advancements in voice-generating AI technology. In recent years, such scams have been on the rise, often targeting vulnerable individuals and resulting in substantial financial losses. AI-driven voice-generating software can replicate a person's voice with a short audio sample, enabling scammers to convincingly mimic trusted voices[40]

7.3 Generated Text

- A new phone application alert highlights the potential misuse of AI in the form of ChatGPT. While ChatGPT has legitimate applications, it can also be exploited by scammers to generate human-like summaries and stories, making it easier to craft well-written scam emails. Criminals can abuse this technology to increase the volume of phishing attacks, potentially leading to users clicking malicious links or divulging personal information. This approach diverges from traditional phishing indicators, which often involve poorly written texts and emails. With the rise of AI technologies aiding content creation, it's increasingly difficult to distinguish phishing attempts from legitimate communication.[41]

7.4 Synthetic IDs

- A woman, known to the immigration authorities as an illegal immigrant utilized two fake passports and IDs to reside and work in a foreign country for seven years. Employing these synthetic IDs, she secured a tenancy in a house, opened bank accounts and obtained employment.[42]

8. INTERPOL's Role in Synthetic Media Forensics

INTERPOL will also continuously work with other stakeholders including experts from the private sector and academia for:

- **Analysis of emerging synthetic media trends, tools, and techniques** to provide critical insights into the evolution of threats and opportunities, enabling proactive and targeted responses.
- **Forensic authentication** to identify methodologies and standards for authenticating media content. These measures will assist investigative efforts, helping to distinguish real content from manipulated material.

[40] https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/
[41] https://abc7chicago.com/what-is-chatgpt-google-chatbot-ai-online-scams/12952645/
[42] https://www.dailymail.co.uk/news/article-12990911/namibian-mother-known-home-office-illegal-immigrant-fake-passport-lived-four-bed-house.html

- **Multi-stakeholder threat assessment** of the criminal landscape to provide an understanding of its potential misuse as well as already evidenced misuse of this emerging technology by criminal actors.

- **Capacity Building and Training** for law enforcement officers on the ever-evolving synthetic media threat landscape and innovative solutions that will help identify, detect and analyze synthetic media files to streamline synthetic media investigations. Since the field of synthetic media forensics is closely related to AI due to the significant role AI plays in both the creation and detection of synthetic media, INTERPOL aims to encourage the responsible use of AI technologies across its member countries.

Reflecting the needs and ambitions of member countries in this area, a dedicated INTERPOL Responsible AI Lab (I-RAIL) has been established. I-RAIL aims to be the focal point for matters related to the responsible use of AI by law enforcement. In addition to facilitating general AI awareness, I-RAIL provides practical support to law enforcement in member countries with regards to theresponsible use of AI, including knowledge development and exchange, agency assessment support and tailored capacity building and training.

9. Conclusion and Recommendations

Synthetic media poses an increasingly formidable challenge to law enforcement agencies worldwide. Its sophisticated manipulation techniques, if left unchecked, can substantially undermine the integrity of evidence and disrupt investigations. To combat this threat effectively, it is crucial for law enforcement agencies to gain a comprehensive understanding of the multifaceted nature of synthetic media, including its creation, distribution, and potential impact on investigations.

Enhancing forensic and investigative capabilities is paramount to developing robust tools and methodologies capable of detecting and differentiating between genuine and manipulated media. Collaborative efforts between countries, industry stakeholders, and academic institutions are essential in pooling resources, sharing expertise, and developing cutting-edge technologies to counter synthetic media misuse effectively. Through ongoing vigilance, continuous technological advancements, and international collaboration, the adverse effects of synthetic media can be mitigated, ensuring the sustained integrity and reliability of law enforcement operations globally.

INTERPOL

INTERPOL | INTERPOL_HQ | @INTERPOL_HQ | INTERPOL HQ | INTERPOL