ITUPublications

International Telecommunication Union
Telecommunication Standardization Sector

Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop

The need for standards collaboration on AI and multimedia authenticity

2024 Report

Disclaimers

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of ITU concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The mention of specific companies or certain manufacturer products does not imply that they are endorsed or recommended by ITU in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by ITU to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader.

The opinions, findings and conclusions expressed in this publication do not necessarily reflect the views of ITU or its membership.

ISBN

978-92-61-39521-6 (Electronic version)
978-92-61-39531-5 (EPUB version)
978-92-61-39541-4 (Mobi version)

Please consider the environment before printing this report.

Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop

The need for standards collaboration on AI and multimedia authenticity

2024 Report

Table of contents

List of figures and tables

Figures

Table

Executive Summary

The International Telecommunication Union (ITU) organized a workshop on "Detecting deepfakes and generative AI: Standards for AI watermarking and multimedia authenticity" on Friday 31 May 2024 during the AI for Good Global Summit. The workshop brought together technology and media companies, artists, international organizations, standards bodies, and academia, to discuss the security risks and challenges of deepfakes and generative artificial intelligence (AI), technological innovations, and areas where standards are needed.

Experts predict that 90 per cent of online content will be generated by AI by 2025. How can we identify whether content was human-generated, AI-generated, or some combination? The problem of AI-generated media and deepfakes is not only technical but also ethical and social. As AI becomes more capable of generating realistic and convincing media content, it becomes harder for humans to discern what is real and what is either completely or partially synthetic. This can erode our trust in the information we receive and the sources we rely on. It can also undermine our sense of reality and identity.

Figure 1: Announcement of new standards collaboration on AI watermarking, multimedia authenticity, and deepfake detection at the AI for Good Global Summit 2024

Some of the key takeaways of the workshop sessions are summarised below:

Challenges presented by generative AI and deepfakes

i) Deepfake and generative AI content is expanding rapidly, in terms of quantity, quality, and the variety of impacts on individuals, organizations, and society at large.

ii) It is generally agreed that deepfake and AI-generated content should adhere to current legislative frameworks that protect copyright or ensure transparency.

iii) Technologies to create deepfake and AI-generated content are becoming more sophisticated and widely available and making it difficult to distinguish between genuine content and synthetic content.

Policies and technical standards for responsible AI use

i) Some governments have introduced specific legislation requiring deepfake and AI-generated content to be labelled, with the aim of preventing AI misuse and ensuring ethical AI development and deployment.

ii) Such legislation should take consumer rights into account, and the burden to identify synthetic content should not lie with the consumer.

iii) Technology and media companies could work towards offering tools for labelling deepfake and AI-generated content on their platforms and make it possible to identify people who post malicious content.

iv) There is a need for standards for multimedia content labelling and authenticity verification and the detection of deepfake and AI-generated content.

Improving deepfake detection

i) Deepfake and generative AI techniques are becoming more sophisticated, making it increasingly difficult to distinguish between content captured by sensors operating in the real world and content synthesized either completely or partially, often with AI.

ii) Technologies for the detection of deepfake and AI-generated content must be continuously updated and upgraded to improve accuracy.

iii) The session on detection technologies highlighted some techniques that can be used to improve detection and performance metrics to benchmark detection technologies.

iv) Detection technologies need to be able to handle various types of data and accurately identify subtle traces and signatures resulting from specific models used to produce deepfake and AI-generated content.

v) There is a need to promote international cooperation and global dialogue on technical standards for detection technologies based on respect for cultural diversity, transparency, safety, and security.

Verifying multimedia provenance and authenticity – *secure metadata, watermarking, and fingerprinting*

i) Tools to establish digital asset provenance and authenticity will be an important part of solutions to the challenge of deepfake and AI-generated multimedia created with malicious intent.

ii) Provenance data for a digital asset can be recorded through Content Credentials based on an open technical standard Coalition for Content Provenance and Authenticity (C2PA). Content Credentials are tamperproof metadata that provide information about the origin, history, and modification of content, including whether the content was AI generated.

iii) Authenticity or provenance verification – the process of assessing content's accuracy and consistency – can help combat misinformation and disinformation and ensure the credibility of multimedia content.

iv) A combination of secure metadata, watermarks, fingerprinting, and secure tools for tracking provenance history is required. C2PA; The Supply Chain Integrity, Transparency, and Trust (SCITT) Working Group of the Internet Engineering Task Force (IETF); and the work of Joint Photographic Experts Group (JPEG) on JPEG Trust provide mechanisms to implement these features.

v) For content credentials to work, they will be needed everywhere – across all devices and platforms – and there will need to be broad awareness of their availability and value.

vi) Standards are needed to enable the interoperability of provenance and authenticity verification mechanisms, calling global collaboration on the development of relevant standards.

A key outcome of the workshop was the decision to set up a multistakeholder standards collaboration for AI watermarking, multimedia authenticity, and deepfake detection convened by ITU under the World Standards Cooperation.

The objectives of the standards collaboration are to:

a) Provide a global forum for dialogue on priority topics for discussion across standards bodies in the area of AI and multimedia authenticity.

b) Map the landscape of technical standards for AI and multimedia authenticity including but not limited to watermarking, provenance, and detection of deepfakes and generative AI content while facilitating sharing of knowledge on lessons learned by different stakeholders.

c) Identify gaps where new standards are required, given the fast-moving nature of the AI and multimedia authenticity landscape.

d) Support the policy, regulatory requirements and government policy measures with regards to AI and multimedia authenticity to facilitate transparency and legal compliance with but not limited to protection of privacy of users, authorship, and the rights of content owners and consumers.

The work in the standards collaboration will be structured under three main areas:

i) Technical Activities – Mapping the standardization landscape for AI watermarking, multimedia authenticity, and deepfake detection with a view to identifying gaps where standards are needed to support related government actions.

ii) Communication – Providing a forum for standards bodies to exchange information and communicate the outcomes of their work.

iii) Policy – Providing a forum for governments and standards bodies to discuss the alignment of policies with standards developed and lessons learned.

Participation in the standards collaboration on AI watermarking, multimedia authenticity, and deepfake detection is open to international, regional and national standards bodies; governments; companies; industry initiatives; and other relevant organizations.

1 Introduction

Experts predict that 90 per cent of online content will be generated by AI by 2025, raising the question of how to identify whether content was human-created, AI-created, deepfaked, or some combination. Deepfakes are a type of synthetic media that can be disseminated with malicious intent. Synthetic media refers to media generated or manipulated using AI. In most cases, synthetic media is generated for gaming and to improve services and quality of life, but the increase in synthetic media and use of generative AI technology has given rise to new possibilities for disinformation and misinformation.

Governments and international organizations are already working towards setting policies, regulations, and codes of conduct to enhance the security and trust of AI systems. The rise of generative AI technology and deepfakes calls for a focus on international standards to support the assessment of multimedia authenticity, the use of watermarking technology, enhanced security protocols, and extensive cybersecurity awareness.

ITU organized a workshop on "Detecting deepfakes and generative AI: Standards for AI watermarking and multimedia authenticity" on Friday 31 May 2024 during the AI for Good Global Summit. The workshop brought together technology and media companies, artists, international organizations, standards bodies, and academia, to discuss the security risks and challenges of deepfakes and generative AI, technological innovations, and areas where standards are needed.

This report outlines key points of discussion at the workshop and the workshop's outcomes, including the recommendation to initiate a multistakeholder standards collaboration on AI watermarking, multimedia authenticity, and deepfake detection.

2 Workshop on AI watermarking and multimedia authenticity

The main objectives of the workshop were to:

a) Provide an overview of the current risks posed by deepfakes and AI-generated multimedia and the related challenges faced by policymakers and regulators.
b) Discuss the effectiveness of AI watermarking, multimedia authenticity, and deepfake detection technologies, their application use cases, governance issues, and gaps that need to be addressed.
c) Discuss the areas where technical standards are required.
d) Explore opportunities for collaboration on standardization activities on AI watermarking, multimedia authenticity, and deepfake detection.
e) Discuss prospective policy measures relevant to global AI governance and their relation to industry-led initiatives such as C2PA and JPEG Trust and the work of international organizations.

The workshop was structured as follows:

i) Session 1: Setting the scene – The challenges and risks of deepfakes and generative AI multimedia.
ii) Session 2: Current state of deepfakes and deepfake detection technology.
iii) Session 3: AI watermarking, multimedia authenticity, and provenance.
iv) Session 4: Standards collaboration to overcome current gaps in AI watermarking and multimedia authenticity.

The workshop considered the various issues related to generative AI and deepfakes, such as:

i) The safety and security risks posed by generative AI and deepfakes.
ii) Deepfake detection technologies and areas where standards are needed.
iii) Multimedia provenance verification, why it is needed, and how it can help address challenges posed by deepfakes.
iv) Policy measures to address deepfakes and generative AI.
v) Areas where standards are needed to support government policies relevant to deepfakes and generative AI.

3 Session 1: Setting the scene – The challenges and risks of deepfakes and generative AI multimedia

The main objectives of this session were to introduce the risks posed by deepfakes and AI-generated content and the policy and legal measures being introduced to address those risks. The panel discussion explored challenges including misinformation and disinformation, consumer protection, copyright protection, and what policy and technological measures would be needed to address them.

Opening Remarks: Bilel Jamoussi, Deputy Director, Telecommunication Standardization Bureau, ITU

Moderator: Alessandra Sala, Senior Director of AI and Data Science, Shutterstock

Speakers

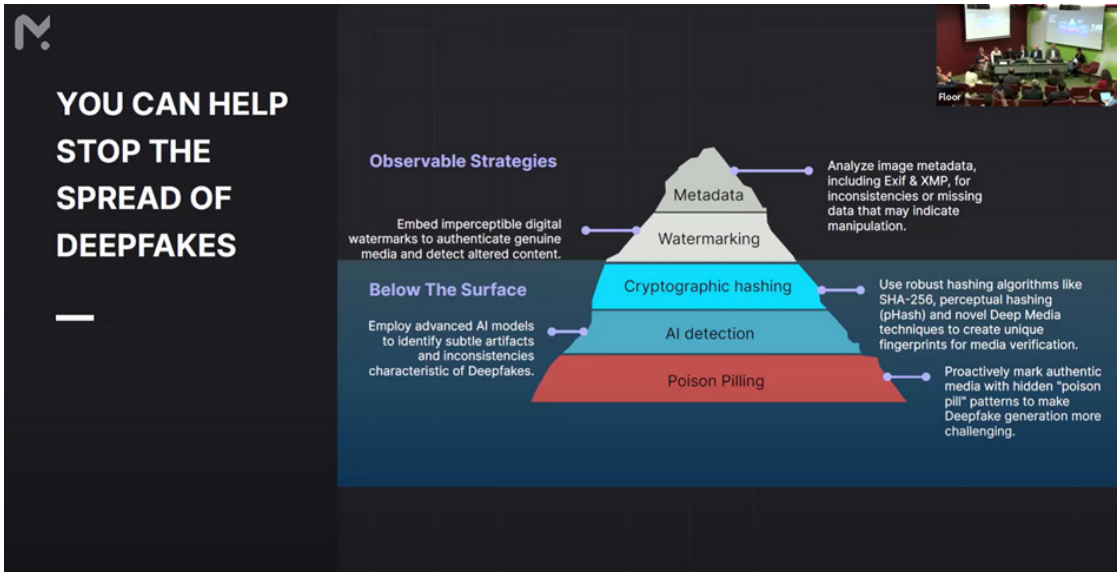
- Sam Gregory, Director, WITNESS.
- Nick Thompson, CEO, The Atlantic.
- Robin Raskin, Founder, Virtual Events Group.
- Harry Yeff, the voice artist Reeps One.
- Helena Leurent, Director General, Consumers International.
- Tobias Bednarz, Legal Counsellor, Copyright Law Division, Copyright and Creative Industries Sector, World Intellectual Property Organization (WIPO).

Bilel Jamoussi, Deputy Director of ITU's Telecommunication Standardization Bureau, provided the opening remarks for the session and highlighted that governments are assigning high priority to the need to address risks posed by generative AI and deepfakes.

The workshop discussed technical solutions and standards already available to address those risks and gaps still to be covered, with the aim of making recommendations on how best to coordinate relevant technical work and provide related information to governments in support of their work on relevant policies and regulations.

Deepfake technology allows people to swap faces in videos and images, change voices, and alter texts in documents. Deceptive videos and images created using generative AI can be used for identity theft, the impersonation of popular figures, disinformation, and the bypassing of identity verification methods to commit fraud.

Figure 2: You can help stop the spread of deepfakes



Source: DeepMedia.AI

The workshop discussions highlighted that tools to generate deepfakes are improving fast, making it increasingly difficult for people to distinguish between an original and a deepfake. In addition, these tools are now widely available in the public domain.

It will become more difficult for journalists and governments to detect deepfakes without easy access to detection tools, said Sam Gregory, Director of WITNESS. Difficulty in trusting the provenance of content on the Internet includes the complicated issue of authentic multimedia content being dismissed as deepfakes. Gregory highlighted that, although the number of observed deepfakes is not yet that high, analysing whether or not content is real or genuine still requires a substantial amount of time and that the results of such analysis then need to be communicated to the public The time required for analysis would become more concerning, therefore, if deepfake generation were to increase substantially.

In addition, Gregory observed that how the content was created, its context, and how it is distributed are perhaps more important than who created the content. The context, for example, could help to determine the motivation for the content's creation and whether it is disinformation or not. With regard to determining the identity of the content creator, various aspects of privacy and data protection would also need to be managed.

Tobias Bednarz, Legal Counsellor at WIPO, spoke to the challenges that generative AI poses in the area of copyright, including questions such as whether or not AI-generated works can be protected by copyright and under what circumstances the generation of content by AI can infringe copyright.

There is broad agreement that content created by an AI system without human intervention is not protected by copyright under current legislative frameworks. AI-assisted works, however, may be protected by copyright.

Addressing the question of under what circumstances the generation of content by AI can infringe copyright – the subject of ongoing lawsuits in several jurisdictions – Bednarz made a distinction between outputs (content generated by AI) and inputs (existing copyright-protected

content used by generative AI models, for example as training data). On the output side, in most jurisdictions, whether or not AI-generated content infringes the copyright held by the creator of a pre-existing work in most jurisdictions will depends on the degree of the AI-generated content's similarity to the copyright-protected work, which can only be assessed on a case-by-case basis. On the input side, copyright holders argue that the unauthorized use of their copyright-protected works to train AI models constitutes copyright infringement. The most important question in this regard is whether such use is covered by limitations to copyright, such as fair use in the US, or exceptions such as those that allow text and datamining, under specific conditions, for the large-scale harvesting of creative works for automated computational analysis. The EU, for example, has a general text and data mining exception, but also allows copyright holders to opt out by reserving their rights in an appropriate manner.

This is also an area in which standardization could be highly beneficial, suggested workshop panellists. Even where opt-out mechanisms exist by law, they do not yet seem operational in practice. Their practical usefulness could increase considerably if international standards enabled copyright holders to control the use of their works in training data efficiently and across AI platforms. Moreover, in the interest of transparency, some jurisdictions such as the EU require providers of AI models to disclose the content they use as training data. Such requirements could be implemented more easily if the way in which this information should be disclosed is further standardized.

Panellists expressed concern about the consequences of misinformation and disinformation on political and social discourse, consequences that could become more pronounced as technologies to create deepfakes become more sophisticated and widely available. Table 1 in Annex 2 summarizes some of the ways that that deepfakes that could affect individuals, organizations, and society at large.

It was highlighted by Helena Leurent, Director General of Consumers International, that legislation being considered by governments to address deepfakes should take consumer rights into account and that the burden of proof should not lie with the consumer. As consumers will not have access to detection tools to identify deepfakes, AI-generated content should be clearly labelled as such.

There was general agreement among panellists that technology companies could work towards offering tools for labelling AI-generated content and make it possible for AI systems to sound less human to help people be certain that they are interacting with an AI system. The recently introduced EU Artificial Intelligence Act contains some policy measures and requirements for transparency aimed at mitigating risks related to AI-generated multimedia and deepfakes (see Annex 1 for more information).

The panellists made clear that there will not be a single solution to the deepfake problem, but rather a combined effort involving technology, legislation, self-regulation, transparency and labelling, standards, education, and incentives for proper usage of AI-generated content.

Key takeaways from this session are summarized below:

i) Deepfakes are expanding rapidly, in terms of quantity, quality, and variety of impacts on individuals, organizations, and society at large.
ii) It is generally agreed that generative AI should adhere to current legislative measures for transparency that support purposes such as the protection of copyright.

iii) Technologies for the creation of deepfakes are becoming more sophisticated and widely available, making it more difficult to identify whether or not content is AI-generated.

iv) Governments have introduced specific legislation requiring AI-generated content to be labelled, with the aim of addressing deepfakes and preventing AI misuse.

v) Legislation to address deepfakes should take consumer rights into account and the burden of proof should not lie with the consumer.

vi) Technology and media companies could work towards offering tools for labelling AI-generated content and make it possible to identify people responsible for deepfakes with malicious intent.

vii) Deepfake is a pejorative term, but it should be recognized that there are circumstances where AI-generated images are desirable, for example, when an actor's appearance is altered to play a character of a different age, or people's appearances are simulated in a virtual environment.

viii) There is a need for standards for multimedia labelling and authenticity verification and the detection of deepfakes.

ix) New systems, often blockchain-based, are being created to provide traceability and digital identity verification.

4 Session 2: Current state of deepfakes and deepfake detection technology

In this session, panellists examined the current state and evolution of deepfake detection technology and examples of innovative technologies used for detecting deepfakes in video, audio, images, and text. Discussions focused on the level of application and performance of detection technologies – and trends in different regions – and their potential to support compliance with relevant policies and regulations planned or already in place.

Moderator: Sam Gregory, Director, WITNESS

Speakers

- Peter Eisert, Professor, Visual Computing, Humboldt University Berlin, Fraunhofer Heinrich Hertz Institute.
- Touradj Ebrahimi, Professor at EPFL, Executive Chairman of RayShaper SA, and Chair of the Joint Photographic Experts Group (JPEG).
- Emma Brown, Co-Founder, DeepMedia, and Rijul Gupta, Co-Founder and CEO, DeepMedia.
- Li Wenyu, Director of Intellectual Property and Innovation Development Center, China Academy of Information and Communications Technology (CAICT).
- Jonghyun Woo, CEO of DualAuth and President of the Passwordless Alliance.
- Wang Ce, Project Manager, China Mobile Research Institute.

Deepfake generation techniques are becoming more sophisticated, making it increasingly difficult to distinguish between real and fake content. This requires that deepfake detection technologies be continuously updated and upgraded to improve accuracy. Deepfaking and deepfake detection will be a long-term and dynamic game.

In his presentation, Peter Eisert from the Fraunhofer Heinrich Hertz Institute discussed the different methods for deepfake generation and detection.

Multiple methods are available to generate deepfakes by manipulating material such as photos, video, and audio are accessible to the public, such as FaceSwap, Jigger, deepfake studio, and so forth. They allow anybody to modify faces in video sequences quickly and simply, resulting in realistic outcomes with little to no effort. In addition, easy access to large-scale public databases and rapid advances in deep learning techniques, notably generative adversarial networks (GAN), have resulted in realistic fake content.

Deepfake detection is typically considered a binary arrangement problem in which classifiers are used to distinguish between reliable and interfering media (videos or photos). This technique requires a big library of real and fake videos or photos to train classification models.

Blind deepfake detection relies on deficiencies or differences in synthetic images. These detection tools rely on information such as:

i) detectable artifacts (e.g., blending, blurring);
ii) inconsistent noise patterns;
iii) temporal inconsistencies (e.g., appearance, geometry, pose, motion);
iv) semantic inconsistencies (e.g., eye blinking, illumination).

Existing deepfake detection tools can be broadly categorized into two groups:

i) Based on either the exploitation of semantic inconsistencies like irregular eye reflections or known generation artifacts in the spatial or frequency domain.
ii) Using neural networks to learn a feature representation in which real images can be distinguished from AI-generated ones. For instance, training a standard convolutional neural network (CNN) on real and fake images from a single GAN yields a classifier capable of detecting images generated by a variety of unknown GANs. Given the rapid evolution of generative AI models, developing detectors which generalize to new generative AI models is crucial and therefore a major field of research.

Deepfake detection techniques use deep learning and machine learning to analyse patterns and anomalies in multimedia content and identify signs of manipulation. Detection techniques can be split into two detection methods: CNN-based methods and region-based convolutional neural networks (R-CNN)-based methods. CNN-based techniques take pictures of people's faces from video frames and feed them into the CNN for training and prediction to get an image-level result. Such algorithms only employ spatial information from a single frame.

R-CNN-based techniques, on the other hand, require a series of video frames for training to produce a video-level result. This approach, known as R-CNN, combines CNN and recurrent neural networks (RNN). As a result, R-CNN-based techniques could fully use spatial and temporal information in deepfake video.

In addition, several deepfake detection methods are based on standard machine learning methods, including utilizing a support vector machine (SVM) as a classifier and extracting handmade characteristics, including biological signals, as a classifier. For example, the video of a person's face contains subtle shifts in colour that result from pulses in blood circulation. These changes in colour form the basis of a technique called photoplethysmography (PPG) that can be used to detect synthetic media. Deepfakes cannot recreate these colour shifts with high fidelity. Biological signals are not coherently preserved in different synthetic facial parts and deepfakes do not contain frames with stable PPG.

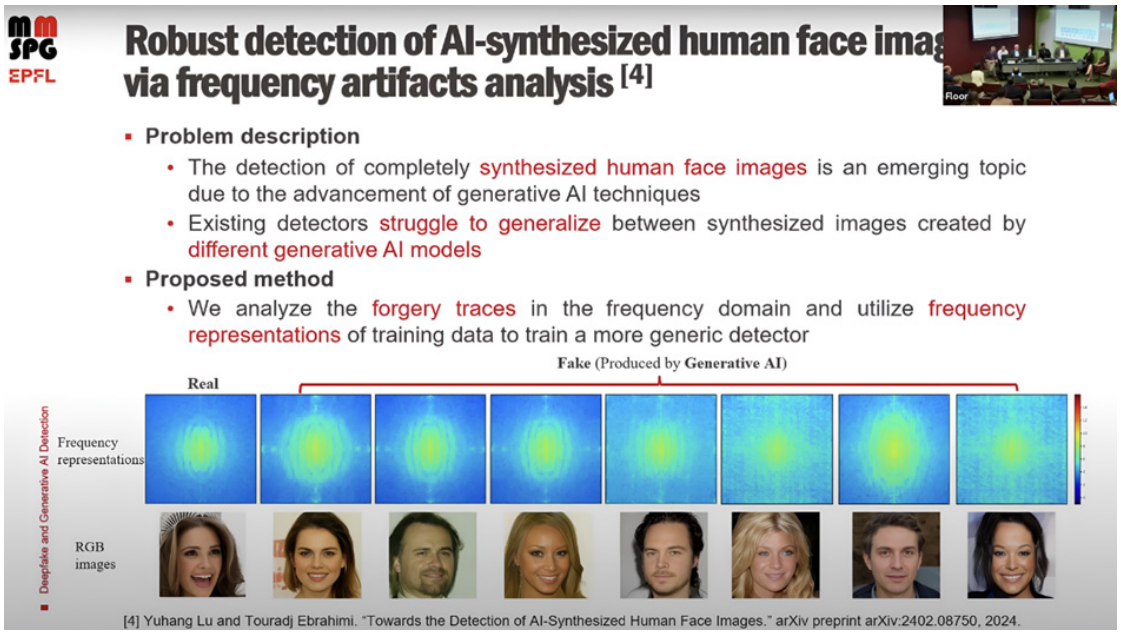
Current deepfake video detection methods have several limitations. Firstly, these methods cannot always be relied upon to detect deepfakes in real-world situations, especially when the images or videos are modified using new techniques that were not present in the training data. Most methods fail to model the natural structures and movements of human faces adequately, which are crucial for accurate deepfake detection. Some methods rely heavily on mouth-related mismatches between auditory and visual modalities, leading to performance degradation when there are limited or unaltered mouth motions. These limitations highlight the need for improved deepfake detection methods that can effectively handle real-world scenarios, generalize well to unseen samples, and capture the natural cues of human faces.

Touradj Ebrahimi, Professor at EPFL and Chair of JPEG, presented a new framework for deepfake detection in still images that could enhance the performance of a deepfake detector under the attack of various real-world perturbations (e.g., JPEG compression artifacts, changing brightness and contrast, blurry effects, and Gaussian noise). He presented two methods:

a) Stochastic degradation-based augmentation.
b) Degradation-based amplitude-phase switch augmentation.

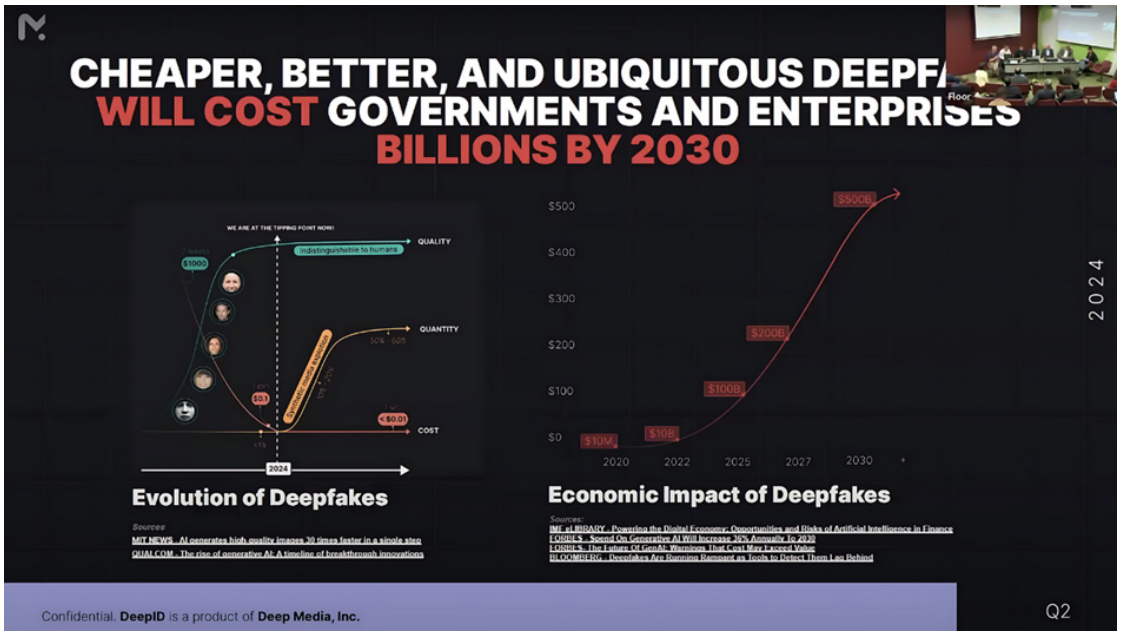
He concluded by presenting a detection technique that allows for the detection of content synthesized completely by generative AI techniques.

Figure 3: Robust detection of AI-synthesized human face images presented by Touradj Ebrahimi, EPFL



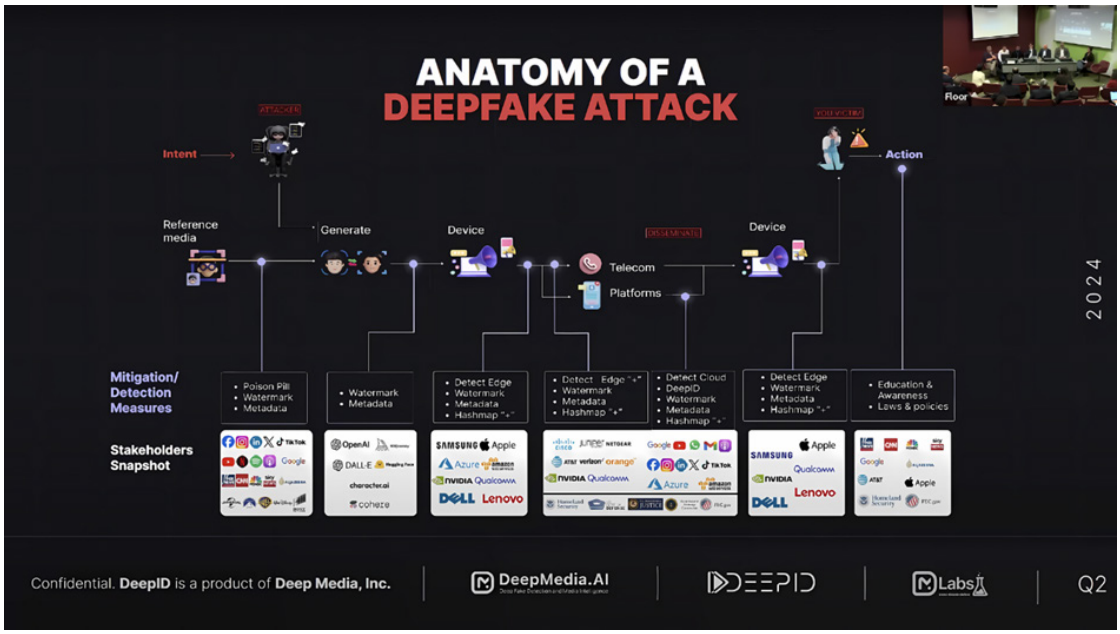
Emma Brown and Rijul Gupta from DeepMedia presented DeepID, a multimodal deepfake detection platform which leverages advanced AI algorithms to analyse multiple data types – including video, audio, and images – and identify manipulated content with high accuracy and speed. Engineered for ease of use, it allows users to adopt and operate the platform without specialized technical expertise. The scalability of DeepID makes it suitable for deployment in various environments – from small-scale operations to large, national-level systems – to safeguard information authenticity on a global scale.

Figure 4: Cheaper, better, and ubiquitous deepfakes will cost governments and enterprises billions by 2023



Source: DeepMedia.AI

Figure 5: Anatomy of a deepfake attack



Source: DeepMedia.AI

Li Wenyu, Director of the Intellectual Property and Innovation Development Centre at CAICT gave a presentation focused on performance evaluation metrics that can help to measure the output of deepfake detection models, ensure the quality and reliability of the models, and further guide the optimisation and improvement of the models to ensure their effectiveness in real-world applications.

Accuracy (ACC), area under the curve (AUC), and average precision (AP) are usually used for assessment:

- ACC is the most intuitive performance metric, reflecting the proportion of samples correctly predicted by the model. It is derived by calculating the ratio of the number of true and true-negative examples to the number of all samples. It is a basic method for assessing how good a classification model is.
- AUC is used to measure the performance of a binary classification system and represent that performance with a value between 0 and 1, with closer to 1 indicating better model performance. It depicts the relationship between the rate of true cases and the rate of false-positive cases at different thresholds.
- AP is an important performance evaluation metric in target detection, which takes into account the accuracy of all the categories of classifiers and averages them. A higher AP means that the model has a better detection accuracy on multiple categories.

The diversity of information forms and the complexity of content today pose a great challenge for deepfake detection. Detection technologies need to be able to handle various types of data and accurately identify subtle traces of a deepfake. There is a need to promote international cooperation and global dialogue on technical standards for deepfake detection technology based on respect for cultural diversity, transparency, safety, and security. The following areas were identified as having potential for standardization in ITU:

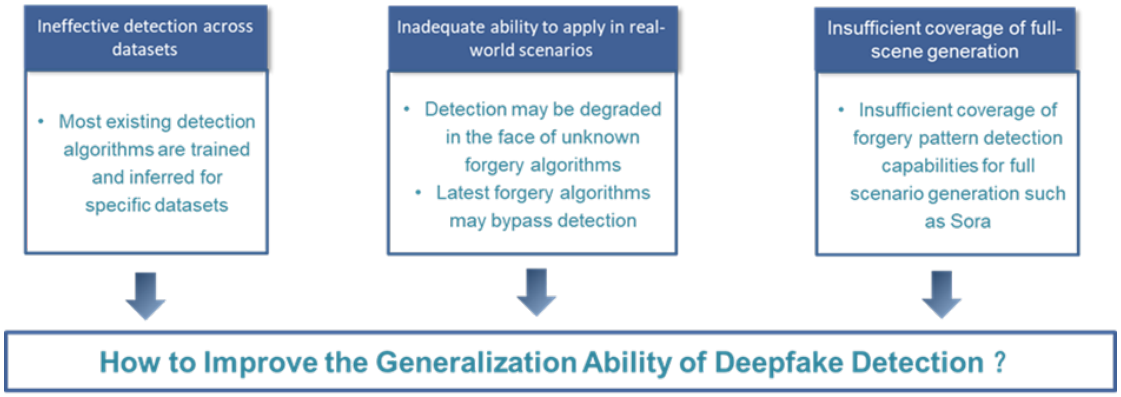
i) Standardization of active defence and traceability – for example, embedding imperceptible watermarks or proof information in multimedia content for content traceability, using blockchain technology to ensure the transparency and tamperproofing of the testing process, and real-time monitoring and tracking using Internet of Things devices.

ii) Standardization of algorithms, models, and other techniques – for example, the standardization of datasets, technologies to detect deepfake and AI-generated content, and ability to adapt to new types of deepfake sand detection techniques.

iii) Standardization of hardware – for example, the hardware acceleration technology of a graphics processing unit or dedicated hardware accelerators that support fast and efficient data processing to achieve real-time detection and the integration and processing of multimodal data, as well as performance benchmarking.

Wang Ce, Project Manager at the China Mobile Research Institute – on behalf of Zhang Chen, CTO of the Security Department of the China Mobile Design Institute – presented the different strategies for improving generalization ability for deepfake detection tools (tools able to detect deepfakes created using different techniques). Some of the strategies presented included data enhancement, adversarial training, adversarial attack, self-supervised learning, multi-task learning, and image reconstruction. With the introduction of text-to-video generative AI models and the development of related technologies, deepfake techniques will not be limited to local replacement and modification, becoming able to synthesize realistic entire scenes. The authenticity of multimedia will be more difficult to discern as a result.

Figure 6: How to improve generalization ability of deepfake detection



Jonghyun Woo, CEO of DualAuth and President of the Passwordless Alliance, gave a presentation focused on how to protect AI systems and training data from deepfakes. The presentation offered an introduction to two ITU international standards, Recommendation ITU-T X.1280 "Framework for out-of-band server authentication using mobile devices" and Recommendation ITU-T X.1220 "Security protection for storage protection against malware attacks on hosts". X.1280 can be used to authenticate the AI system that the user is connecting to and X.1220 can protect the training data model from unauthorised access, tampering, or malware affecting the integrity of the training dataset.

5 Session 3: AI watermarking, multimedia authenticity, and provenance

The main objective of this session was to focus on multimedia provenance and provide examples of industry initiatives and areas where standards are needed. This session examined industry-led initiatives such as C2PA's Content Credentials specification. The session also considered how a combination of AI watermarking to identify AI-generated multimedia and Content Credentials tied to a real-world identity via cryptography could help in protecting copyright.

Moderator: Emma Brown, DeepMedia

Speakers

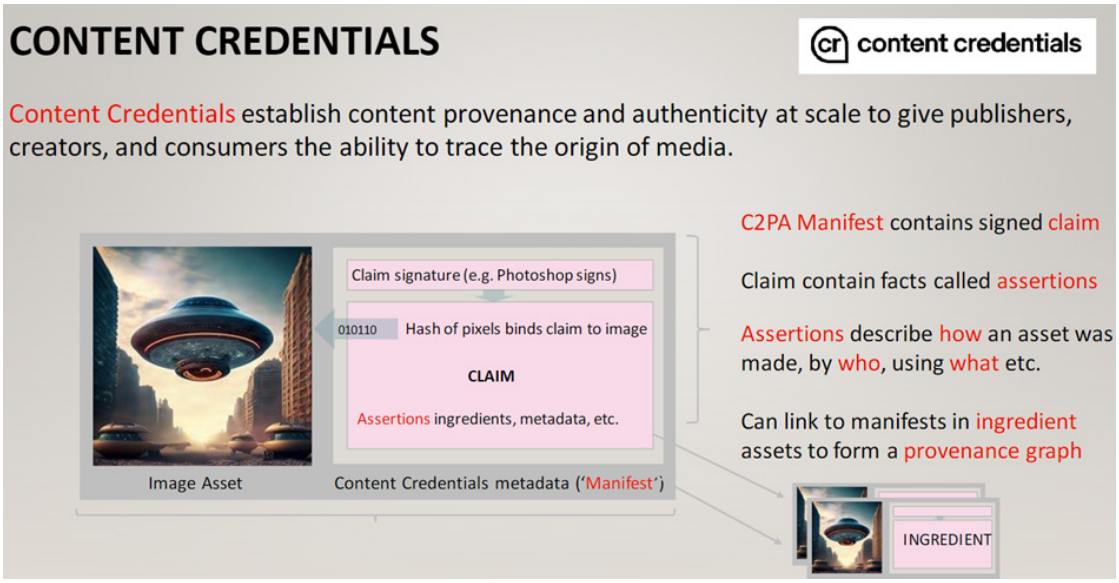
- Leonard Rosenthol, Chief Architect, Adobe.
- Andrew Jenks, Director of Media Provenance, Microsoft, and Chair of C2PA.
- Hu Zhengkun, Director of AI Ethics and Governance Research, SenseTime.
- Jon Geater, Chief Product and Technology Officer, DataTrails, and Co-Chair of the IETF SCITT Working Group.
- Touradj Ebrahimi, Professor at EPFL, Executive Chairman of RayShaper SA, and Chair of JPEG.

Provenance – the record of AI-generated content's origin and history – is crucial for verifying authenticity, protecting creators' rights, and enhancing content quality.

Authentication or provenance verification – the process of assessing content's accuracy and consistency – can help combat disinformation and ensure the credibility of multimedia content. They can be used to enhance transparency and accountability in various industries by providing a secure and transparent way to track the provenance and ownership of digital assets.

Leonard Rosenthol, Chief Architect at Adobe, provided an overview of C2PA's Content Credentials specification. Andrew Jenks, Director of Media Provenance at Microsoft and Chair of C2PA, explained that the specification is an example of affirmative authenticity which represents a cryptographically secure method of marking content.

Figure 7: Content Credentials



Source: Leonard Rosenthol presentation

Content Credentials are based on C2PA's open technical standard and used to establish digital asset provenance and authenticity. The specification is designed to support the emergence of a global ecosystem of provenance-establishing applications to meet the needs of a wide range of individuals and organizations. The specification is also designed to meet appropriate security and privacy requirements and help uphold human rights.

Figure 8: Visuals generated by various AI software 6 months ago, presented by Andrew Jenkins

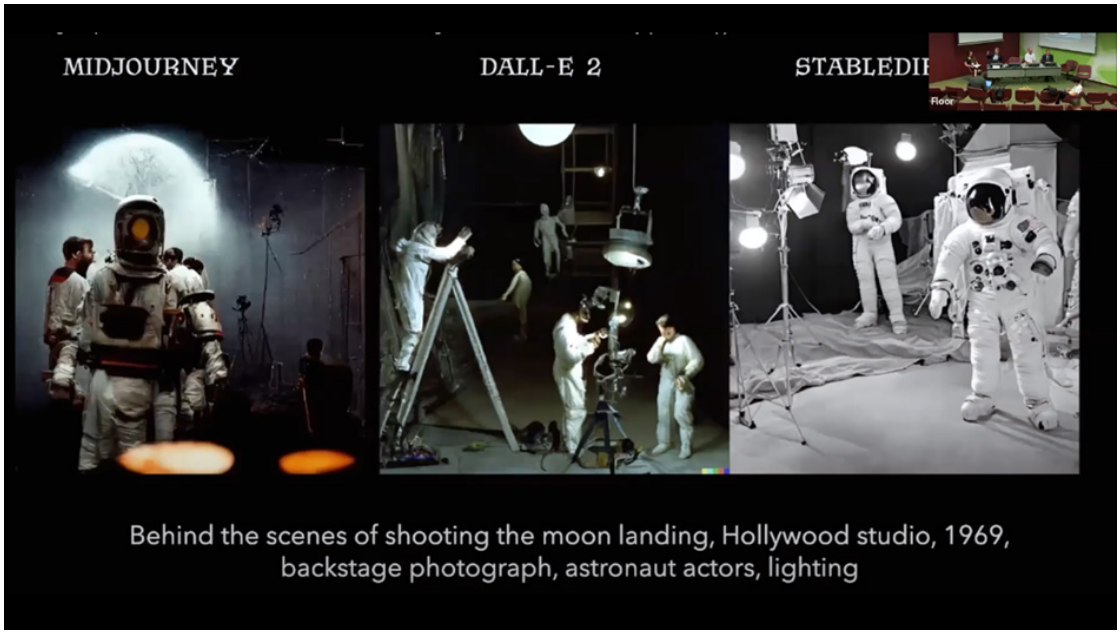
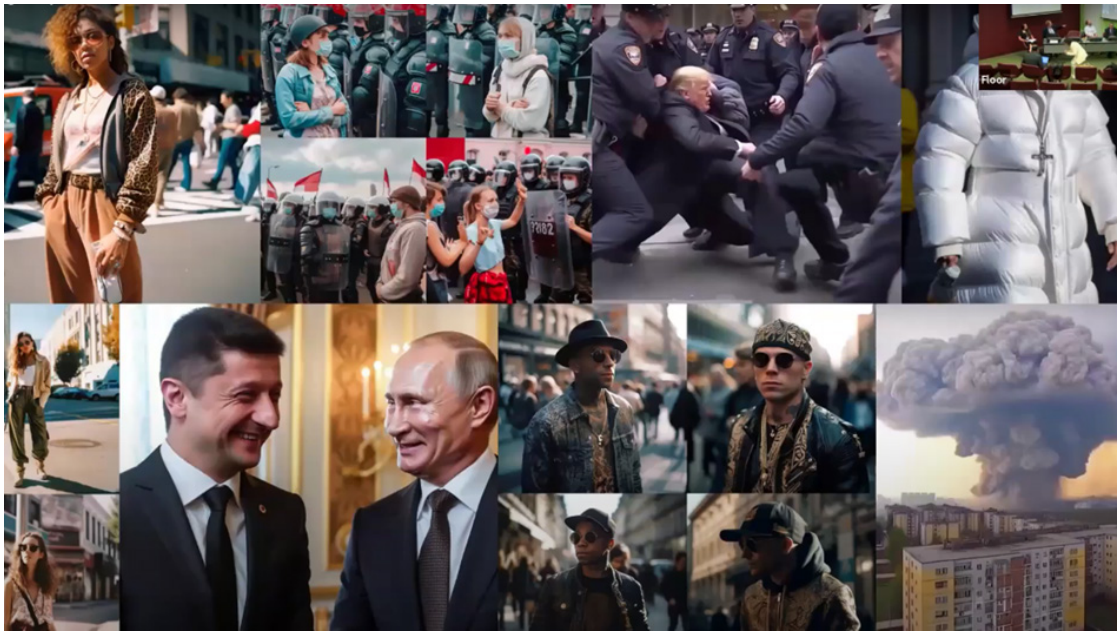


Figure 9: Mosaic of AI-generated images shown during the workshop by Andrew Jenks to illustrate the rapid advancement of AI in contrast with Figure 3. Warning: these images are not real.



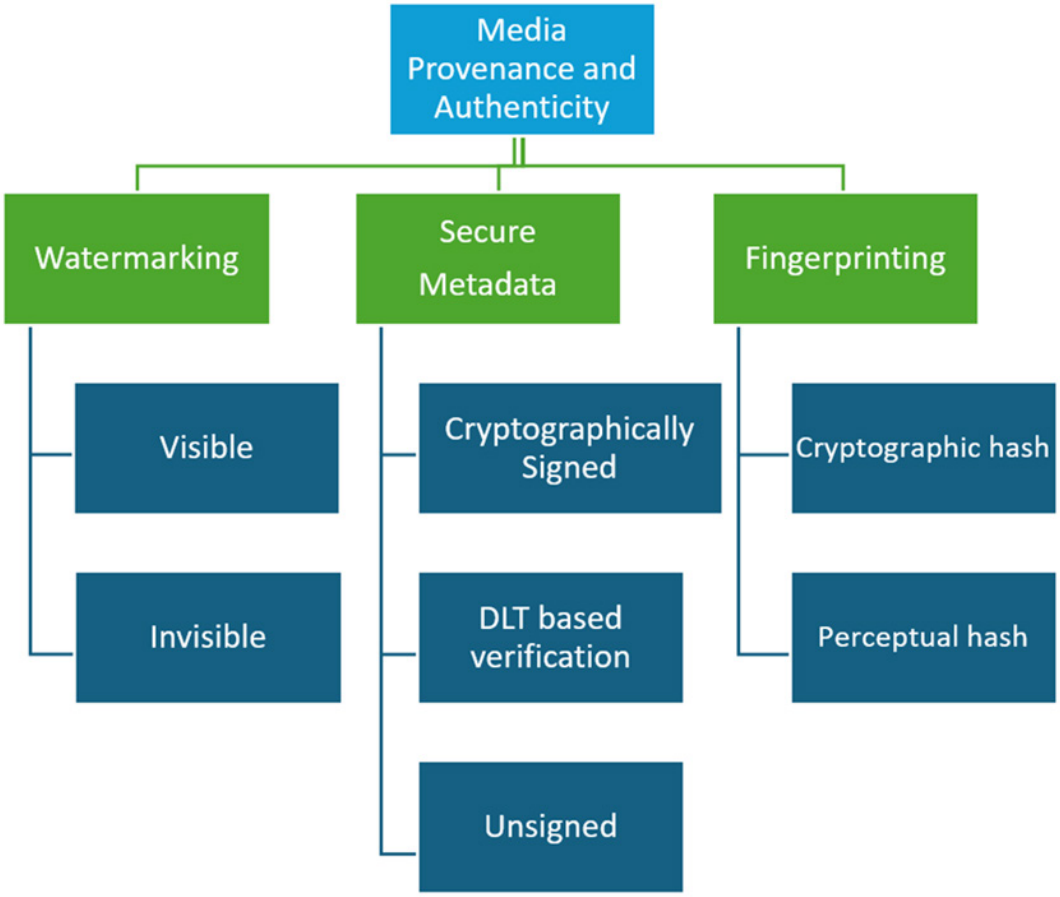
The main advantage of Content Credentials is that they can combine secure metadata, watermarks, and content fingerprinting to provide a solution to establish the provenance of multimedia content.

Some of the techniques for recording and preserving provenance data include:

- **Secure metadata**: This is verifiable information about how content was made associated with the content itself, in a way that cannot be altered without leaving evidence of alteration. Content Credentials can provide information about the provenance of any piece of multimedia content or composite. It indicates whether a video, image, or sound file was created with AI or captured in the real world with a device like a camera or audio recorder. Content Credentials are designed to be chained together to provide information about how content may have been altered, what was combined to produce the final content, and even what device or software was involved in each stage. The various elements of provenance in the lifecycle of a digital asset can be combined in ways that preserve privacy and enable creators and consumers to obtain information on the origins of the digital asset and how it was created. The metadata can be either signed using digital signatures or preserved using distributed ledger technologies such as blockchain to maintain the integrity of the provenance data and verify its authenticity.
- **Watermarking**: Watermarking is a technique to embed hidden information that cannot be seen by people. It can be decoded using a watermark detector. State-of-the-art watermarks can be impervious to alterations such as the cropping or rotating of images or the addition of noise to video and audio. Importantly, the strength of a watermark is that it can survive efforts to remove secure metadata such as taking screenshots or pictures of pictures, or re-recording audiovisual content.
- **Fingerprinting**: This refers to a way to create a unique code or hash based on pixels, frames, or audio waveforms that can be computed and matched against other instances of the same content, even if there has been some degree of alteration. The fingerprint can be stored separately from the content as part of Content Credentials. When someone encounters the content, the fingerprint can be re-computed and matched against a database of Content Credentials and its associated stored fingerprints. The advantage

of this technique is that it does not require the embedding of any information in the media itself. It is immune to information removal because there is no information to remove.

Figure 10: Techniques for establishing media provenance and authenticity



None of these techniques are durable enough in isolation to be effective on their own. However, when combined, the three techniques form a unified solution that is robust and secure enough to ensure that reliable provenance information is available.

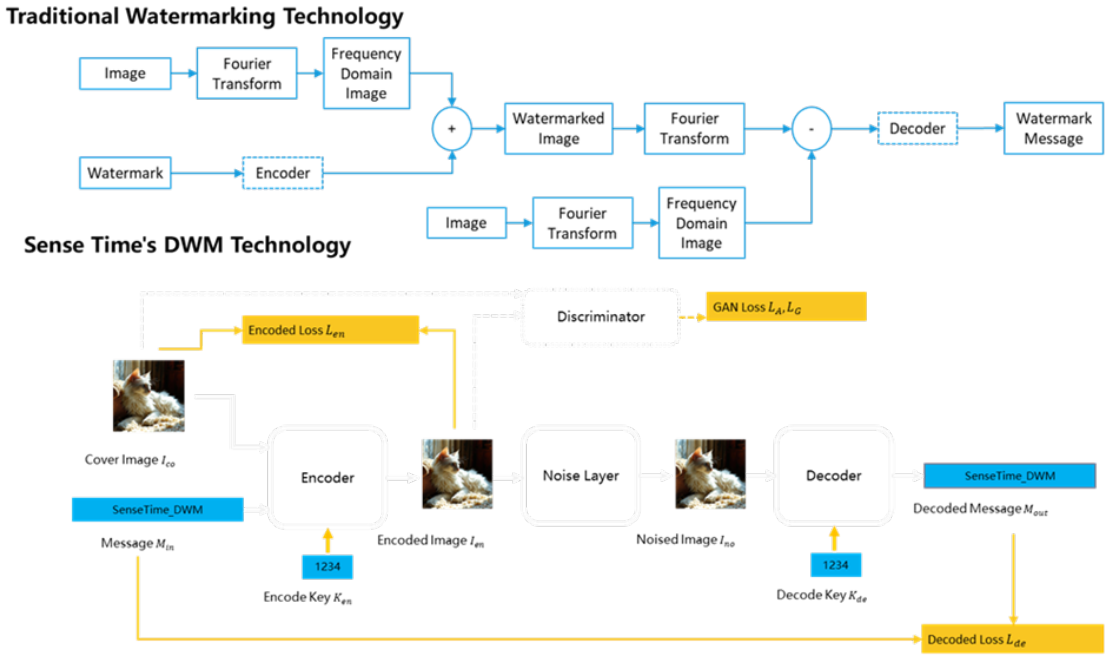
Although Content Credentials show potential to be an important part of the ultimate solution for the problem of deepfakes, they will not form the entire solution. For Content Credentials to work, they will be needed everywhere – across all devices and platforms – and there will need to be broad awareness of their availability and value. Moreover, there is an urgent need for industry-wide standards and protocols for establishing content provenance and authenticity to enable global interoperability.

Hu Zhengkun, Director of AI Ethics and Governance Research at SenseTime, presented SenseTime's solution for establishing content provenance and authenticity based on digital watermarking in compliance with government policy requirements in China on content labelling and the transparency of generative AI. The digital watermarking solution features:

- Watermarking methods based on deep neural networks, which achieve better robustness and adaptive capability as well as generalization ability than traditional frequency domain watermarking methods.
- Encryption algorithms and keys for the security of the encoding and decoding process.

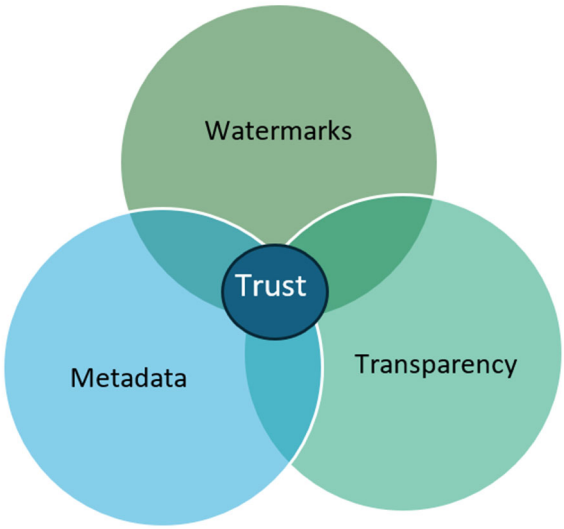
- Robustness with the extraction of watermarking not being affected by moderate cropping, cropping fill, pixel replacement, image compression, zooming, rotation, noise, blurring, combination, and other interferences.
- Invisible watermark that currently supports the watermarking of bitmap images, vector graphics, videos, and audio.

Figure 11: Comparison between SenseTime's solution and traditional watermarking methods



Jon Geater, Chief Product and Technology Officer at DataTrails and Co-Chair of the IETF SCITT Working Group, mentioned in his presentation that all work on relevant regulations and standards is pointing towards durable, high-integrity accountability, transparency, and data provenance for AI-generated content.

Figure 12: Trust elements



Jon Geater presented the SCITT framework being discussed IETF. This framework provides a multi-layered approach that combines a few core technologies to meet most challenges:

i) Portable metadata helps create a lingua franca to describe and understand the provenance of multimedia content across different platforms.
ii) Watermarks help to identify images, even if altered. They are also harder to remove than metadata and thus more durable identifiers.
iii) Transparency technology helps to hold actors accountable and settle disputes. Advantages include:

    a) Broad applicability.
    b) Independent storage of provenance metadata with the ability to reunite content with its history if that history has been removed.
    c) Flexible, resilient cryptographic proofs.
    d) Prevents shredding and backdating, even with multiple copies of files in multiple systems.

Touradj Ebrahimi, Professor at EPFL at Chair of JPEG, provided an overview of JPEG Trust, a new international standard planned for publication this year. The JPEG Trust framework addresses aspects of authenticity, provenance, and integrity through secure and reliable annotation of images throughout their lifecycle. The JPEG Trust framework will provide building blocks for more elaborate use cases and it is expected that the standard will evolve over time and be extended with additional specifications.

Some of the key takeaways of this session were:

i) Tools to establish digital asset provenance and authenticity will be an important part of solutions to the challenge of deepfake and AI-generated multimedia created with malicious intent.
ii) Provenance data for a digital asset can be recorded through Content Credentials which are tamperproof metadata that provide information about the origin, history, and editing process of the content (including whether or not it was AI-generated).
iii) Authentication or provenance verification – the process of assessing content's accuracy and consistency – can help combat misinformation and disinformation and ensure the credibility of multimedia content.
iv) A combination of secure metadata, watermarks, fingerprinting, and secure tools for tracking provenance history is required. C2PA, SCITT, and JPEG Trust provide mechanisms to implement these features.
v) For Content Credentials to work, they will be needed everywhere – across all devices and platforms – and there will need to be broad awareness of their availability and value. vi. Standards are needed to enable the interoperability of provenance and authenticity verification mechanisms, calling for global collaboration on the development of relevant standards.

6 Session 4: Standards collaboration to overcome current gaps in AI watermarking and multimedia authenticity

The session concluded the workshop by highlighting the main issues discussed and identifying a way forward for global collaboration on the development of technical standards for AI watermarking, multimedia authenticity, and deepfake detection that will consider the requirements of both developed and developing economies to support safe, sustainable, inclusive, and trustworthy generative AI.

Moderator: Alessandra Sala, Senior Director of AI and Data Science, Shutterstock

Speakers

- Bilel Jamoussi, Deputy Director of the ITU Telecommunication Standardization Bureau.
- Silvio Dulinsky, Deputy Secretary-General, International Organization for Standardization.
- Gilles Thonet, Deputy Secretary-General, International Electrotechnical Commission.
- Thomas Wiegand, Fraunhofer Heinrich Hertz Institute.
- Andrew Jenks, Director of Media Provenance, Microsoft, and Chair of C2PA.
- Leonard Rosenthol, Chief Architect, Adobe.
- Jon Geater, Chief Product and Technology Officer, DataTrails, and CoChair of the IETF SCITT Working Group.

Figure 13: Discussion on Standards Collaboration in Session 4

Governments are already working towards setting policies, regulations, and codes of conduct to address the challenges related to deepfakes and generative AI. Technical standards will play an important role in supporting the policies and regulations being introduced by governments. This was highlighted during discussions at both AI Governance Day and the workshop covered by this report at the AI for Good Global Summit 2024.

To make progress as an industry as a whole, it was agreed to set up a multistakeholder standards collaboration for AI watermarking, multimedia authenticity, and deepfake detection convened by ITU under the World Standards Cooperation.

The objectives of the standards collaboration are to:

a) Provide a global forum for dialogue on priority topics for discussion across standards bodies in the area of AI and multimedia authenticity.
b) Map the landscape of technical standards for AI and multimedia authenticity including but not limited to watermarking, provenance, and detection of deepfakes and generative AI content while facilitating sharing of knowledge on lessons learned by different stakeholders.
c) Identify gaps where new standards are required, given the fast-moving nature of the AI and multimedia authenticity landscape.
d) Support the policy, regulatory requirements and government policy measures with regards to AI and multimedia authenticity to facilitate transparency and legal compliance with but not limited to protection of privacy of users, authorship, and the rights of content owners and consumers.

The work in the standards collaboration will be structured under three main areas:

i) Technical Activities – Mapping the standardization landscape for AI watermarking, multimedia authenticity, and deepfake detection with a view to identifying gaps where standards are needed to support related government actions.
ii) Communication – Providing a forum for standards bodies to exchange information and communicate the outcomes of their work.
iii) Policy – Providing a forum for governments and standards bodies to discuss the alignment of policies with standards developed and lessons learned.

Participation in the standards collaboration on AI watermarking, multimedia authenticity, and deepfake detection is open to international, regional and national standards bodies; governments; companies; industry initiatives; and other relevant organizations.

Annex 1: EU AI Act

Some governments have introduced specific legislation to address deepfakes and generative AI, prevent misuse and ensure ethical AI development and deployment. Deepfakes are addressed specifically under the EU AI Act due to their potential risks. The main provisions regarding deepfakes include:

1) Transparency Obligations:
    - Developers and users of deepfake technologies are required to clearly disclose that the content is AI-generated. This is aimed at preventing misinformation and ensuring that audiences are aware of the artificial nature of the content they are viewing. It also ensures recognition of the authors whose works were used in the creation of AI.
    - Labelling of AI content is mandatory/suggested by classification and watermarking of deepfakes.
2) High-Risk Classification:
    - Deepfakes used in contexts that can significantly impact individuals' rights or society (e.g., political manipulation, defamation) may be classified as high-risk and thus subject to stricter regulatory requirements.
3) Accountability and Traceability:
    - Traceability and accountability in the creation and dissemination of deepfakes is to be ensured. This involves maintaining records of the processes and data used to generate deepfakes, enabling authorities to track their origins if necessary.
4) Prohibited Uses:
    - Certain malicious uses of deepfakes, such as those intended for social scoring or illegal surveillance, are prohibited under the Act's unacceptable risk category.

The AI Act also clarifies that general-purpose AI models need to put in place a policy to comply with EU copyright law.

The EU's Code of Practice on Disinformation addresses deepfakes through fines of up to 6 percent of global revenue for violators. The code was initially introduced as a voluntary self-regulatory instrument in 2018 but now has the backing of the Digital Services Act. The Digital Services Act, which came into force in November 2022, increases the monitoring of digital platforms for various kinds of misuse. Under the EU AI Act, deepfake providers would be subject to transparency and disclosure requirements.

Annex 2: Impact of Deepfakes

Table 1: Impact of Deepfakes

Who is impacted?	Types of impact		
	Reputation	Financial	Dis – misinformation and manipulation of information
Individual	Intimidation Defamation Swapping of women faces in pornographic images/videos	Identity theft Phishing Extortion Financial scams	Attacks on politicians Cyberbullying
Organizational	Brand reputation Undermining trust	Stock price manipulation Fraud	Fabricated court evidence Media manipulation Fake news
Society wide	News media manipulation Impact on economic stability Erosion of trust Damage to democracy Manipulation of elections Damage to international relations Damage to national security		

International Telecommunication Union
Place des Nations
CH-1211 Geneva 20
Switzerland

ISBN: 978-92-61-39521-6

Published in Switzerland
Geneva, 2024

Photo credits: ITU