POLICY PAPER: BUILDING TRUST IN MULTIMEDIA AUTHENTICITY THROUGH INTERNATIONAL STANDARDS

CAROL BUTTLE AND CINDY PAROKKIL

POLICY PAPER
Edition 1

TABLE OF CONTENTS

AI AND MULTIMEDIA STANDARDS COLLABORATION

The AI and Multimedia Authenticity Standards Collaboration is a global initiative advancing standardization in the rapidly evolving field of AI-generated and altered media. By identifying gaps and driving the development of new standards, we support transparent, privacy-conscious, and rights-respecting practices. Our work also aims at informing policy and regulatory frameworks to promote legal compliance and safeguard public trust.

Led by the World Standards Cooperation[1], the collaboration serves as a vital forum for dialogue among standards developers, civil society organizations, technology companies, and other key players. Participating organizations include the International Electrotechnical Commission (IEC), the International Organization for Standardization (ISO), the International Telecommunication Union (ITU), the Coalition for Content Provenance and Authenticity (C2PA), the China Academy of Information and Communications Technology (CAICT), DataTrails, Deep Media, and Witness.

Convened by ITU under the auspices of the World Standards Cooperation, the collaboration was launched at the AI for Good Global Summit in 2024.

Learn more here (https://aiforgood.itu.int/multimedia-authenticity/) or contact the Secretariat at amas-secretariat@itu.int

[1] International Electrotechnical Commission (IEC), the International Organization for Standardization (ISO), and the International Telecommunication Union (ITU)

PREFACE

This paper is primarily aimed at policymakers and regulators. It seeks to demystify the complexities of regulating the creation, use and dissemination of synthetic multimedia content through prevention, detection and response, and to present these issues in a clear and accessible manner for audiences with varying levels of expertise and technical understanding. In addition, the paper aims to highlight global initiatives and underscore the vital role and benefits of international standards in promoting regulatory coherence, alignment and effective enforcement across jurisdictions.

The document offers practical guidance and actionable recommendations, including a regulatory options matrix designed to help policymakers and regulators determine what to regulate (scope), how to regulate (voluntary or mandatory mechanisms), and to what extent (level of effort). It also explores a range of supporting tools – such as standards, conformity assessment mechanisms, and enabling technologies – that can contribute to addressing the challenges of misinformation and disinformation arising from the misuse of multimedia content. At the same time, it emphasizes the importance of striking a balance that enables the positive and legitimate use of either fully or partially synthetic multimedia for societal, governmental and commercial benefit.

Finally, the paper includes a set of practical checklists for use by policymakers, regulators and technology providers. These can be used when designing regulations or enforcement frameworks, developing technological solutions or preparing crisis response strategies. The checklists are intended to help align stakeholder expectations, identify critical gaps, support responsible innovation, and enable conformity with emerging standards and best practices.

ACKNOWLEDGEMENT

This paper was developed under the Policy Pillar of AMAS, led by ISO, and co-authored by Carol Buttle and Cindy Parokkil. We gratefully acknowledge the valuable contributions of AMAS members to the development of this policy paper. Any errors that remain are entirely the authors' own responsibility.

Section **01**

THE CONTEXT

Generative Artificial Intelligence (GenAI) has the potential to be one of the most transformative technologies seen for decades. To realize its potential, however, requires not only recognizing the immense benefits it offers but also acknowledging and managing the significant risks it involves. Historically, the societal impact of emerging technologies has depended on the speed, breadth and depth of their adoption. In the case of GenAI, adoption has accelerated at an unprecedented pace, raising the stakes for thoughtful design and robust governance.

As GenAI becomes increasingly integrated throughout all sectors, there is a growing need for comprehensive frameworks encompassing policy, regulation, standards, and compliance and certification. These frameworks must embed safeguards and ethical principles into GenAI systems from inception and design. This presents a formidable challenge for policymakers, particularly in the face of fragmented global legal and regulatory landscapes, as they navigate the complexities of a technology that carries a potent transformative power to transform society and economies throughout the developed and developing world alike. The urgency for international coordination has never been greater.

> **“Generative AI’s potential impact and risks transcend national borders, demanding a global scope for new policy and technology solutions.”**
>
> The Organisation for Economic Co-operation and Development

The Organisation for Economic Co-operation and Development (OECD) has emphasized that GenAI must be understood through a global lens, with policy and technical solutions developed accordingly. Unlike the industrial revolutions of the 18th and 19th centuries, which began in the UK before spreading to Europe and the US, the AI revolution is global and simultaneous. Countries must now navigate the dual challenge of realizing GenAI’s benefits in domains such as governance, healthcare, defence and civil society, while mitigating risks and protecting citizens from misuse.

Synthetic media – any content that is generated or manipulated using artificial intelligence (AI), such as deepfakes, AI-generated text, images or voice – presents both opportunities and serious challenges.

Beyond the obvious proliferation of misinformation and disinformation, there are issues of erosion of trust; as synthetic media becomes more realistic, it becomes harder to distinguish real from fake. Furthermore, with many countries lacking clear laws about the creation and use of synthetic media, legal and ethical ambiguities arise, raising questions about consent, ownership and accountability.

In a world where digital identity is increasingly important and prevalent, the risk of identity theft and fraud has grown too. Synthetic identities are commonly created to commit financial fraud or manipulate digital identity systems. Biometric security systems are also prone to attacks from facial deepfakes and voice cloning.

GenAI and synthetic media offer multiple opportunities, but these are only achievable if supported by a comprehensive policy that focuses on transparency, disclosure and harm mitigation.

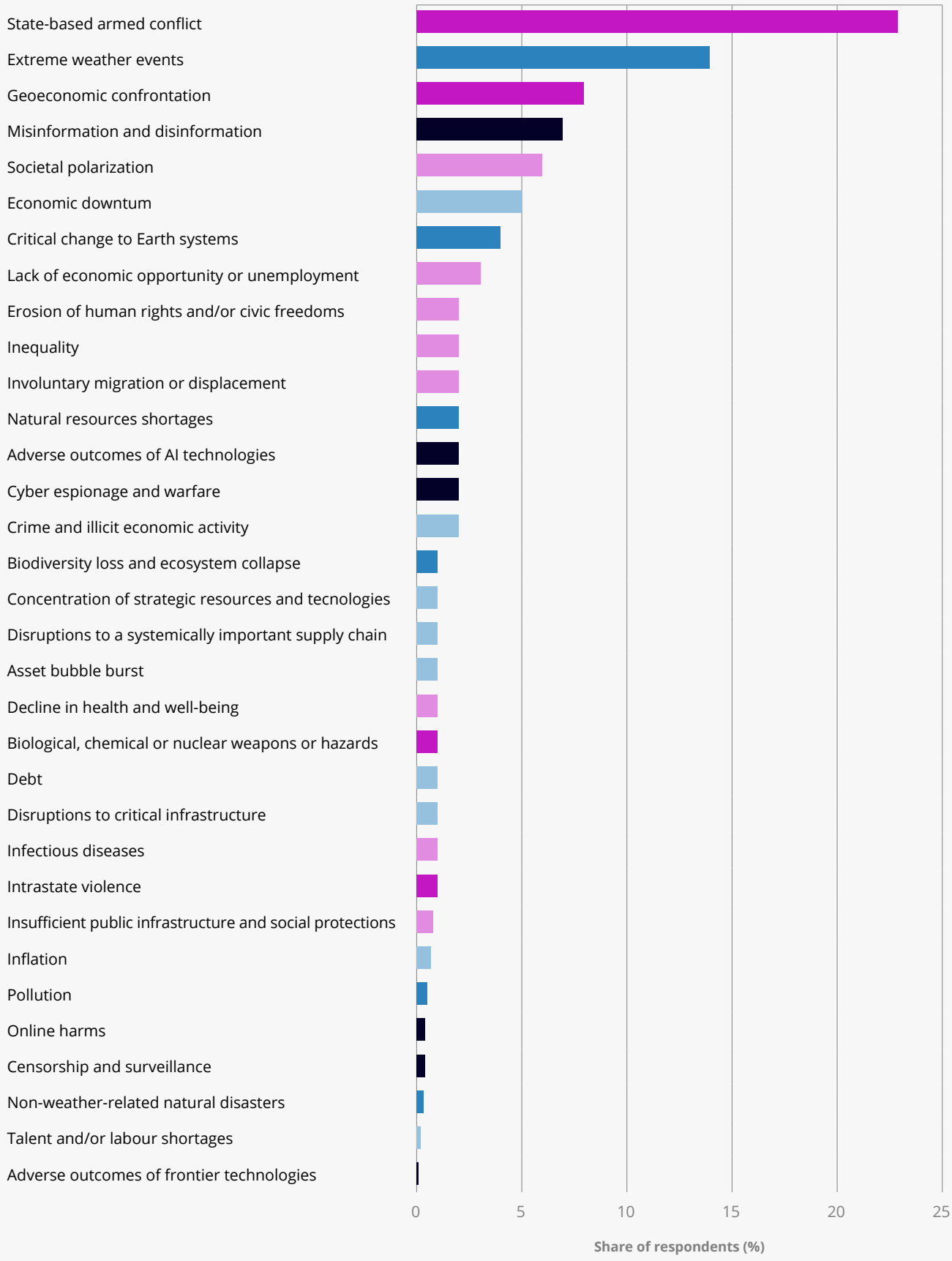
1.1 Misinformation and disinformation in the age of AI

Concerns about the authenticity of information have existed for centuries. However, the digital age and environment – accelerated by AI – has magnified these issues, turning them into global, cross-border threats with significant implications for public trust, national security and democratic institutions.

The scale, speed and sophistication of digital content creation and dissemination have outpaced traditional methods of content verification. New tools and strategies are required to validate content, protect intellectual property and preserve public trust without stifling innovation.

These challenges and their impact on society have rapidly escalated the issue of misinformation and disinformation to the level of public policy. Governments worldwide are responding with a mix of regulatory instruments, technical standards and public awareness campaigns.

Misinformation and disinformation are now ranked among the world's most pressing risks. According to the World Economic Forum's "Global Risks Report 2025", misinformation and disinformation remain the top global risk for the second consecutive year. The growing sophistication of GenAI-generated content makes it increasingly difficult to discern truth from falsehood, particularly as synthetic media blurs the line between real and fabricated experiences.



Source: World Economic Forum Global Risks Report 2025[2]

1.2 Definitions matter: Misinformation, disinformation and malinformation

Misinformation and disinformation have become almost interchangeable terms, but they are distinct from one another, especially in their motives and application. Often overlooked in discussions is malinformation. Malinformation, in the context of fake news, can be especially dangerous when used in conjunction with disinformation as part of orchestrated campaigns intended to spread untruths.

- Misinformation refers to false information but is not created or shared with the intention of causing harm.[3]
- Disinformation is false content intentionally created and disseminated to mislead, harm or manipulate.
- Malinformation is factual information used out of context with the intent to cause harm. For example, publishing private data with malicious intent (e.g. revenge porn or non-consensual intimate imagery), or altering contextual metadata to mislead.

A table of different types of misinformation and disinformation has been provided in Annex 1.

> **“There are many ways in which a proliferation of false or misleading content is complicating the geopolitical environment. It is a leading mechanism for foreign entities to affect voter intentions; it can sow doubt among the general public worldwide about what is happening in conflict zones; or can be used to tarnish the image of products or services from another country.”**
>
> World Economic Forum

Tactics such as propaganda, scams and fake news are not new, but digital technologies have made them more accessible, scalable and potent. Historically used as tools of war and politics, disinformation today can be deployed by state and non-state actors alike, with devastating consequences for vulnerable populations such as refugees, migrants and marginalized communities.

[2] https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf

[3] https://webarchive.unesco.org/web/20230926213448/https://en.unesco.org/fightfakenews, or non-consensual

In today's hyperconnected digital environments, disinformation behaves much like a contagion – its rapid spread threatens to destabilize public discourse and erode democratic resilience. When false narratives are systematically deployed – whether by domestic actors or foreign entities – they can undermine public trust in critical areas such as healthcare, climate policy and national security. These campaigns cast doubt on empirical evidence, deepen societal divisions, and make it harder to form the collective consensus needed to address complex global challenges.

1.3 Who are the types of perpetrators presenting challenges?

Several types of actors are targeted to spread misinformation and disinformation

Individuals: Ordinary citizens can intentionally (or even unintentionally) spread harmful content, knowingly or unknowingly. Technologies such as deepfakes make it easier than ever to fabricate convincing images and audio.

Political candidates and organizations: Candidates and political entities may exploit false narratives to influence public opinion, deepen polarization and undermine electoral integrity.

Social media platforms: These platforms prioritize engagement over accuracy, enabling the viral spread of falsehoods. Echo chambers reinforce existing beliefs, making corrections harder to reach those affected.

Model developers and providers: These give rise to multiple challenges that range from content authenticity and attribution, the amplification of misinformation, a lack of transparency on how models are trained, the data they use and how those outputs are moderated.

Legacy media: Legacy outlets are not immune to manipulation, especially in the digital age, despite traditional safeguards. Deepfakes and unmoderated user content (e.g. comment sections) further complicate the issue.

Nation-states/foreign actors: Nation states and foreign actors may use coordinated disinformation strategies – such as troll farms or sponsored influencers – as part of Foreign Malign Influence Subversive operations to destabilize societies and manipulate public opinion.

These activities have a significant financial and social cost, as illustrated in Figure 1

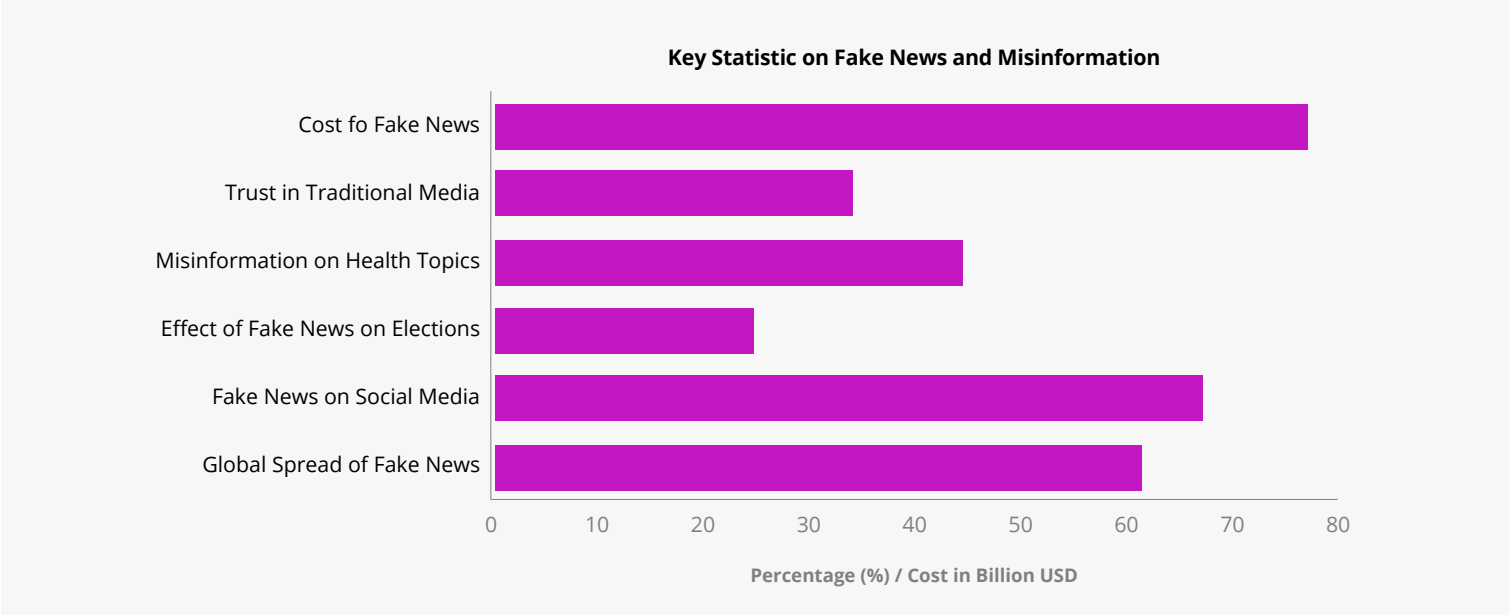


Figure 1: Key statistics on Fake News and Misinformation[4], Source SDLC Corp

1.4 Deepfakes and cyber-attacks

Deepfakes,[5] initially created for entertainment and artistic purposes, are now being weaponized. The ease with which adversaries fabricate realistic images, videos and audio recordings, and the growing inability to distinguish between synthetic and non-synthetic content is providing cybercriminals with ample opportunity to launch sophisticated attacks.

"Deepfakes and the misuse of synthetic content pose a clear, present, and evolving threat to the public across national security, law enforcement, financial, and societal domains."[6]

Department of Homeland Security, United States

Hyper-realistic images, videos and audio recordings are increasingly used in sophisticated fraud, identity theft and social engineering attacks.

[4] Source: https://sdlccorp.com/post/fighting-fake-news-how-blockchai n-can-verify-media-authenticity/

[5] See Annex 3 for categories of deepfakes

[6] https://www.govtech.com/artificial-intelligence/wyoming-lawmakers-grapple-with-ai-regulation-debate

The financial sector is especially vulnerable. A recent Medius report found that 53 % of finance professionals had been targeted by deepfake scams, with 43 % falling victim.[7] In one notable case, a finance employee was tricked into transferring $39 million to fraudsters using deepfake video.[8]

The consequences go beyond financial loss. Public figures – including politicians, celebrities and influencers – face significant reputational damage. Ironically, the very media that sustains their careers can be manipulated against them.

1.5 AI and multimedia authenticity

Trust in digital content/multimedia is built on the belief that its integrity, origin, lineage and context are preserved. This includes confirmation that creators of any of those mediums follow strict ethical practices that avoid plagiarism, misinformation and disinformation. When content is altered without consent – especially in legal, financial or journalistic settings – the ramifications can be significant.

For organizations, ensuring content lifecycle integrity (from creation through to management and distribution) is increasingly difficult. Questions arise over who created, modified or consumed a piece of content, and whether it still reflects the truth. Failure to meet basic standards in quality and content governance exposes individuals and institutions to legal and regulatory (including data protection and intellectual property rights), and reputational risk.

Forgery and media manipulation have long existed, from forged paintings to altering photographs. The difference today is the scale and speed with which GenAI can replicate, fake or distort reality. For example, spirit photography in the 1800s or doctored portraits of Abraham Lincoln pale in comparison with today's deepfakes, which can impersonate voices and identities with frightening precision.

This not only endangers victims but erodes public confidence in all forms of media, leading to outright dismissal of authentic media. This has profound implications for journalism, governance, justice and social cohesion, especially if legitimate evidence is wrongly perceived as fabricated. The risk is in no longer just being fooled, it's in becoming cynical of everything, including the truth.

As GenAI blurs the line between synthetic and non-synthetic, it becomes harder for individuals to trust what they see, hear or read. This crisis of credibility affects everyone from governments to businesses, journalists, educators and the public. Different groups will experience different levels of impact based on exposure and vulnerability (see Annex 2 for stakeholder impacts). As trust in content declines, the risks span throughout legal, social, ethical and technical domains. As the spread of misinformation grows and credibility of sources declines, we face a complex challenge that spans technical, ethical and social concerns. What is needed is urgent innovation in content verification, combined with greater digital literacy, which can be supported by sound legal and regulatory frameworks and international standards.

[7] https://www.medius.com/media/vqfj0a0b/medius-financial-census-2024.pdf

[8] https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam

Section **02**

THE COMPLEXITIES OF BALANCING THE REGULATORY LANDSCAPE WITH MARKET NEEDS TO BUILD TRUST IN MULTIMEDIA AUTHENTICITY

In 2013, the World Economic Forum identified the "rapid spread of misinformation online" as one of the top 10 global risks. More than a decade later, this concern remains at the forefront. In its "Global Risks Report 2025", the organization reaffirmed that misinformation and disinformation are among the world's most pressing challenges.

> **"Misinformation and disinformation remain top short-term risks for the second consecutive year, underlining their persistent threat to societal cohesion and governance by eroding trust and exacerbating divisions within and between nations."**
>
> World Economic Forum

Despite repeated warnings and growing financial, societal and reputational consequences, the question remains: Why does the challenge persist?

2.1 Why is building trust in multimedia authenticity complex?

Building trust in multimedia authenticity is inherently challenging due to the interdependent components of its ecosystem and the wide range of stakeholders involved. Compounding this issue is the absence of a globally accepted digital identity framework, which makes it difficult to reliably validate the identity of individuals or organizations, particularly across borders. As a result, the landscape is increasingly vulnerable to identity theft, impersonation and synthetic identities.

Achieving trust requires the following:

- Clear international and national policies and regulations that establish a comprehensive and coherent framework,
- Organizational compliance and support throughout sectors to consistently apply these frameworks,
- Technological solutions that are designed and deployed in line with regulatory requirements, and
- Robust enforcement mechanisms, both mandatory and voluntary, to ensure consistent and meaningful implementation.

2.2 What factors contribute to the difficulties?

There are several factors contributing to the difficulty of achieving this:

- It is a global issue, but implementation and enforcement occur nationally (or sometimes even at the local level), often influenced and shaped by varying political philosophies and jurisdictional constraints as well as market requirements. For example, it is critical to reach agreement on penalties for non-compliance and enforcement action across borders.
- The issue cuts across multiple sectors and domains – including consumer protection, intellectual property and national security – meaning no single regulation can address the full scope of multimedia authenticity. The EU's General Data Protection Regulation (GDPR), although focused on data privacy, provides a worthy basis for other areas to follow. The GDPR has extraterritorial effect, despite its focus on the EU and UK, and as a result has initiated a deliberation of similar laws in other countries facing similar issues.
- Policymakers need to balance competing priorities, such as preventing online harms while protecting freedom of expression, encouraging innovation and attracting investment.
- Successful implementation of regulation depends on strong support and collaboration from industry, including the development of compliant technological solutions.
- Levels of regulatory capacity and maturity vary. Countries differ significantly in their ability to develop, implement and enforce regulations, making global alignment and coordination a major challenge.
- There is a tension between the rapid pace of technology and the lag in regulation. The rapid evolution of GenAI, cloud computing and cross-border data flows outpaces regulatory systems. Jurisdictional ambiguity over data residency and the lack of a central global internet authority further exacerbate fragmentation.

2.3 Overview of regulatory landscape

A combination of global principles, international guidelines, and national regulatory frameworks are increasingly guiding efforts to regulate online safety, misinformation and disinformation. These involve contributions from governments, international organizations, civil society and technology companies, often emphasizing concepts such as safety-by-design, transparency and accountability.

Given that misinformation spans sectors and platforms, a diverse range of legal and policy mechanisms have emerged globally. Below is an overview of key international initiatives, followed by regional and national regulatory frameworks.

International initiatives

- OECD Digital Service Providers Guidelines: These promote a risk-based approach, particularly to protect children and vulnerable users.
- Global Internet Forum to Counter Terrorism (GIFCT): This is a public-private partnership designed to detect and remove harmful content.
- The Christchurch Call: Led by New Zealand and France, this initiative brings together governments and tech companies to eliminate terrorist and violent extremist content.
- GPAI and OECD Initiatives: This promotes the responsible use of AI in content moderation and information integrity.
- UNESCO Guidelines for Regulating Digital Platforms (2023). These outlines human rights-based principles to address misinformation and disinformation, complementing the UN Guiding Principles on Business and Human Rights.
- UNESCO's Recommendation on the Ethics of Artificial Intelligence: Adopted in 2021, this is applicable to all 194 UNESCO member states and makes recommendations for policy action areas.
- UNESCO's Media and Information Literacy (MIL) Framework: This is designed to empower users to critically assess the reliability of information and enhances digital literacy globally.

The Global Online Safety Regulators Network, established in 2022, is a coalition of international online safety regulators (including Australia's eSafety Commissioner, UK's Ofcom, Ireland's Coimisiún na Meán, Fiji's Online Safety Commission, South Korea's Broadcasting and Communications Commission (BCC) and others from Europe, North America and Asia-Pacific). It aims to:

- Promote safe online environments through cooperation,
- Support evidence-based policy development, and
- Encourage alignment in regulatory approaches without enforcing one-size-fits-all solutions.

Its Online Safety Regulatory Index[9] provides a comparative analysis of how different jurisdictions approach online safety regulation and provides:

- National legislative models,
- Enforcement maturity,
- Common principles (e.g. child protection, systemic risk), and
- Global trends and convergence/divergence in practice.

It helps policymakers track global trends, aids platforms with compliance across jurisdictions, and promotes interoperability among regulations.

[9] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/international/other/global-online-safety-regulators-network-regulatory-index.pdf?v=383839

Regional and national approaches and frameworks

European Union:

- General Data Protection Regulation (GDPR), 2018
  - Focused on privacy, it also restricts data misuse and algorithmic profiling that can fuel misinformation (e.g. microtargeting).
- Digital Services Act (DSA), 2022
  - Comprehensive regulation for online platforms,
  - Mandates content moderation transparency, algorithmic accountability and mitigation of systemic risks like disinformation,
  - Applies stricter rules to Very Large Online Platforms (VLOPs), and
  - Mandates rapid response to disinformation and hate speech.
- EU Code of Practice on Disinformation, revised in 2022
  - A voluntary but increasingly institutionalized code signed by major platforms, including Meta, Google, etc.
  - Requires transparency in political advertisements, demonetization of false content and support for fact-checkers.
- Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law provides recommendations on:
  - Fact-checking,
  - Platform-design solutions, and
- Empowerment of users.
- EU AI Act
  To address deepfakes, the EU's AI Act promotes transparency with Article 50(2). It requires providers of general-purpose AI tools to tag AI-generated content and identify manipulations, enabling users to better understand the information. However, this does not apply to standard editing tasks like minor corrections, or where authorized, for law enforcement activities like crime detection or prosecution.

  The EU AI Act, particularly Recital 133, acknowledges the need for flexibility to accommodate various content formats, detection methods and AI functionalities. This ensures efficient compliance for providers, especially those dealing with diverse content and evolving technologies. Recital 133 further emphasizes the importance of accurate, compatible, and effective tools for tagging and identification, including technologies like watermarks, metadata tags, fingerprints or security features to trace content origin and prove authenticity. A key concern involves its ambiguity regarding deepfake classification. While it requires disclosure of AI-generated content, the EU AI Act avoids explicitly designating deceptive deepfakes as high risk.

Africa:

- African Union Convention on Cybersecurity and Personal Data Protection (Malabo Convention), 2014
  - Encourages African Union member states to enact laws against cybercrime, with protections for data privacy and freedom of expression.
- National examples (e.g. Nigeria, Kenya and South Africa):
- Often use cybercrime and hate speech laws to address disinformation, although concerns over freedom of speech persist.

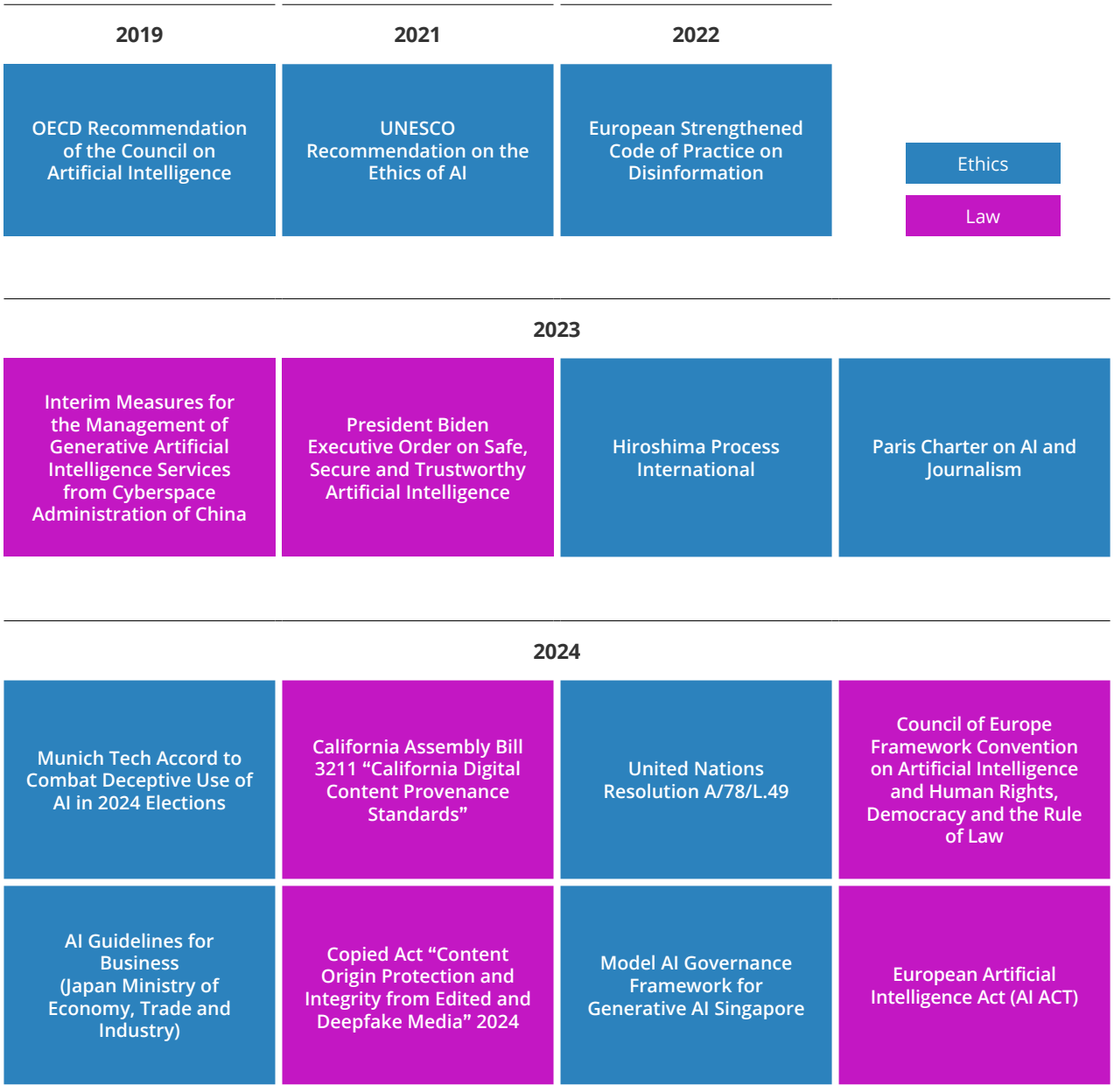
Asia-Pacific:

- ASEAN Digital Masterplan 2025 promotes digital safety cooperation and media literacy throughout South-East Asia.
- Australia: Online Safety Act, 2021, empowers the eSafety Commissioner to remove harmful content. Emphasizes safety-by-design and protects against cyberbullying and misinformation.
- India: IT Rules (2021) requires swift content takedown, traceability of originators, and imposes stricter rules on 'significant platforms'.
- Singapore: Protection from Online Falsehoods and Manipulation Act (POFMA), 2019, allows government-issued correction orders or blocking access to false content. It faces criticism over potential free speech impacts.
- China: Provisions on the Administration of Deep Synthesis Internet Information Services (2023) regulates the use of GenAI and deepfake technologies. It requires platforms to label AI-generated content, prevent misuse and ensure synthetic media does not spread false or harmful information.

Americas:

- United States: No comprehensive federal law on misinformation due to First Amendment protections. Key elements include:
  - Section 230 of the Communications Decency Act: Provides platform immunity while enabling moderation,
  - FTC enforcement: This targets deceptive commercial practices related to disinformation, and
  - State-level efforts: e.g. California Age-Appropriate Design Code Act addresses children's safety.
  - Canada:
  - Online Harms Act (Bill C-63, 2024, proposed): This aims to regulate harmful online content, including hate speech and misinformation, and
  - Digital Citizen Initiative: This funds education and research combatting disinformation.
- Brazil: Fake News Bill (PL 2630, proposed): This seeks to mandate user ID verification, track viral messages and disclose sponsored content, particularly to combat electoral and health misinformation.
- United Kingdom Online Safety Act, 2023: This imposes duties of care on platforms to address illegal and harmful content, especially affecting children. Regulated by Ofcom, it includes misinformation provisions with broad social impact.

International standards, in conjunction with initiatives and collaborations, form a powerful mechanism for achieving regulatory collaboration, and are crucial to building user trust and enabling safe deployment of AI-powered multimedia technologies. The following visual shows how the progression of ethical and legal frameworks are developing for content labelling:



2.4 Bridging the gap between regulation and trust

One of the major challenges faced by policymakers and regulators is that multimedia authenticity, like GenAI, is fundamentally a 'Black Box', particularly in the context of online safety regulation. There is limited transparency about how these models are developed and trained. Technologies offer significant potential for good, but the question that looms is how to enable effective governance when the underlying operations are largely opaque. The main challenges about how to ensure trustworthiness and interpretability of multimedia content without stifling innovation intersects with broader concerns. These include how to align with emerging global priorities, such as combatting misinformation, and how they can be shaped or influenced by online safety regulation.

The Global Online Safety Regulators Network in their first Annual Report[10] and Strategic Plan for 2025-2027[11] have highlighted the following themes as focus points:

- Building regulatory coherence across jurisdictions,
- Contributing to the evidence base of online safety and surfacing best practices, and
- Facilitating the sharing of information and coordination to promote compliance.

There is currently confusion and a lack of clarity about the status and application of key online safety measures and the type of online harms they address. This has a major bearing on enabling risk mitigation in relation to misinformation and disinformation. By the very nature of a technology that exploits a lack of borders, without visibility of one region's approach, a position of equitable and recognisable governance will be difficult to enforce. Definitive understanding of the territorial scope of regulations, how different jurisdictions are mobilizing standards and laws, and their status as presented above is both a current challenge and one that will continue.

Working out how to achieve a framework of agreed policy and regulation based on applicable and appropriate international standards, and one that can be future-proofed in a way that allows it to advance in line with technology, is a vast problem that requires multistakeholder collaboration. Online safety measures are an integral part of the overall fight against all areas of multimedia usage and the harms that can ensue. Without coordination we risk allowing a gap in approach that will be difficult to retrospectively close.

Yet, the disparity between differing nation's approaches to misinformation, disinformation, deepfakes and multimedia authenticity can be bridged, and cohesion can be achieved. No one underestimates the size of the task and there are many collaborative projects ongoing with a common mission to find solutions that bear witness to the sheer effort required.

[10] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/international/other/gosrn-annual-report-2024.pdf?v=386966

[11] https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/international/other/gosrn-three-year-strategic-plan-publication-2025-to-27.pdf?v=386967

What these collaborative projects show is a recognition from multiple stakeholders that regulatory and enforcement bodies cannot build trust in multimedia alone. We need all parties to work together and find new forms of international collaboration and regulation, even perhaps self-regulation. This needs to be coupled with corporate responsibility that fosters trust and includes human rights, media literacy and ethics of the individual user.

Yet, calling for parties to work together and promoting initiatives will remain a largely philosophical trend if we constantly debate the issues without developing solutions that are workable and can be applied.

As we discuss in the next section, one way of developing these initiatives is to propose practical solutions that build on existing frameworks and standards and can be adopted by governments and industry.

2.5 Finding practical solutions for governments and industry

Many of the challenges highlighted above can be better understood and addressed by examining how different governments are increasingly adopting Prevent-Detect-Respond (PDR) frameworks to build trust in multimedia authenticity. This three-pronged approach provides a scalable, flexible structure that balances regulatory intent with technical feasibility.

Table 1. Applying PDR framework to MMA

Approach	Policy Requirements	Method	Benefit and/or outcome
Prevention	Transparency	Labelling	Informs users about various aspects of the content. Clearly identifying if the content was AI generated.
		Watermarking	Non-human perceptible markings applied to content that provide information about it.
	Traceability	Content provenance tools	Enables providing information about the content's origin and changes to establish accountability and attribution.
	Accountability	Conduct risk assessment	Enforcement can be made more efficient when areas are identified as high risk. Prevalent abuse or patterns of behaviour are identified and treated as priorities. This proactive approach helps mitigate the risks associated with manipulated content, ensuring that users are protected from misinformation and fraudulent activities.
	User education	Public awareness initiatives	Reduces accidental misuse through education about copyright laws and the consequences of infringement.

Approach	Policy Requirements	Method	Benefit and/or outcome
Detection	Detecting manipulated content and deepfakes	Technological solutions	These solutions offer numerous benefits such as protecting intellectual property, verifying image, audio, text and video authenticity, and aiding in online safety and security. However, it creates a 'back and forth war' with bad actors who attempt to avoid these detectors. For example: https://arxiv.org/abs/2504.2148
	Data privacy	Data handling and adherence to data protection legislations	All data processed are subject to randomized manual review, ensuring accuracy and compliance with data protection legislation.
Response	Enforcement	Regulatory interventions	Penalties can be applied and rules enforced through governments enacting laws and regulations that specifically address the techniques and approaches that should be used. They also address what happens when such techniques are breached.
	Explainability	Use of explainer-type algorithms, AI model verification methods and information about training datasets used.	Decisions made by AI systems can be checked to maintain a high level of reliability and trustworthiness. This helps mitigate risks of IPR breaches.
	Dispute mechanisms	Content contestability	Clear and well communicated mechanisms benefit individuals, helping them dispute claims.
		Platform bans	Policing of problematic areas can be more effective and beneficial when access to platforms and websites that frequently host infringing content is regularly removed.

This framework mirrors successful approaches in privacy (e.g. GDPR, CCPA) and cybersecurity (e.g. NIST cybersecurity framework,[12] PCI-DSS). The strength of PDR lies in its simplicity and versatility; it is widely understood, adaptable throughout sectors, and conducive to regulatory alignment. In the case of privacy, successful approaches emphasize prevention (privacy-by-design), detection (breach notification and monitoring), and response (enforcement actions and mechanisms for user redress). These regulations appear to primarily focus on privacy, but they offer a valuable model for tackling multimedia authenticity by highlighting the importance of clear, transparent and accurate information in how data – particularly partially or fully synthetic content – is used and communicated.

[12] NIST's Cyber Security Framework expands on Prevent, Detect and Respond with additional functions of Identify and Recover. https://www.nist.gov/cyberframework/getting-started/online-learning/five-functions

However, applying a PDR framework requires more than a technical lens; it demands a socio-technical perspective. This involves recognizing the complex interplay between human behaviours and ethical use, business processes and market incentives, and technology design and deployment, within each phase of prevention, detection and response.

When implemented at the organizational level, PDR-based frameworks increase the likelihood of achieving regulatory alignment, consistency and equitable compliance. This common structure helps foster adoption, encourage accountability and streamline communication between governments and market actors.

Moreover, PDR enhances the enforceability of regulations. When both public and private sectors operate with a common structure, regulatory goals become more actionable. This is precisely where international standards and conformity assessments play a critical role in implementing PDR.

The next section explores the role of international standards in bridging the policy-technology gap, and outlines specific standards that can support trust in multimedia authenticity. Ultimately, the effectiveness of any PDR framework – especially in complex domains like multimedia authenticity – relies on how well it is aligned with relevant international standards. These standards provide the technical and procedural foundations necessary to support each of the PDR pillars.

By grounding future regulation in proven models like PDR and embedding standards at every level, stakeholders can collectively create a more trustworthy and resilient digital information ecosystem.

Section **03**

THE ROLE OF INTERNATIONAL STANDARDS AND CONFORMITY ASSESSMENT IN ADDRESSING MULTIMEDIA AUTHENTICITY

The rapid evolution of GenAI, its growing influence on multimedia creation and editing and/or manipulation, as well as the increasing spread of misinformation and disinformation, pose increasingly complex challenges for governments and regulators worldwide. To effectively address these risks, coordinated and harmonized action is essential, particularly in the development of standards and specifications that enable mutual recognition of mechanisms for verifying multimedia authenticity. Without this, cross-border regulatory gaps will persist, leading to fragmentation, inefficiencies and vulnerabilities.

International collaboration is the cornerstone of an effective response. International standards, conformity assessment procedures, and the broader Quality Infrastructure (QI) system should underpin this collaboration.[13] These tools not only provide the necessary technical and governance frameworks to meet today's challenges, but also ensure regulations evolve in tandem with rapid technological developments.

To promote mutual recognition of content authenticity and close cross-border loopholes, governments should adopt and reference internationally recognized, consensus-based standards. Among other things, international standards offer policymakers:

- A shared vocabulary and set of common benchmarks that support interoperability across jurisdictions,
- Evaluation methods and best practice frameworks for safety, security, governance and accountability, and
- A mechanism to avoid technological 'lock-in' or 'lock-out' by promoting open, flexible and adaptable solutions.

Without alignment with international standards and clear agreement on how conformity assessment can be used, the risk of further regulatory fragmentation, duplication of effort, and inefficient allocation of public and private resources will only increase. It is to be noted that not all standards are created equal. Priority should be given to internationally agreed standards developed through transparent, multistakeholder processes.

[13] As defined by INetQI: A system that includes organizations (public and private), policies, legal frameworks, and practices needed to support the quality, safety, and environmental soundness of goods, services, and processes. It's a comprehensive framework that underpins the functioning of markets and facilitates access to foreign markets.

3.1 The value of international standards

International Standards, as developed by ISO, IEC and ITU, jointly known as the World Standards Cooperation (WSC) are global tools that respond to market needs and reflect the consensus of diverse global experts. Developed through inclusive, multistakeholder processes, these standards address social, environmental, technical and economic dimensions.

The OECD's Good Regulatory Practices and the World Trade Organization's Technical Barriers to Trade (WTO TBT) Agreement both advocate for the use of international standards in regulation. These standards are aligned with the WTO TBT's six principles for the development of international instruments, meaning they are presumed not to create unnecessary obstacles to trade and enable regulatory cooperation. When referenced in regulations, policies or conformity assessment schemes, international standards can:

- Reduce regulatory burden by providing ready-made best practices,
- Accelerate policy implementation by separating technical rules from political cycles, and
- Facilitate international cooperation and smooth global trade flows.

In the context of public policy, particularly in advancing the United Nations Sustainable Development Goals (SDGs), international standards enhance transparency, predictability and accountability. They offer a cost-effective, efficient means of implementing policy while fostering sustainable economic growth.

Contrary to the common misconception that standards hinder innovation, a growing body of research demonstrates that well-developed international standards support and drive innovation. They provide stable foundations for research and development, promote interoperability, and reduce duplication of effort, enabling innovators to focus on delivering differentiated, value-added solutions. This is particularly vital in fast-paced, competitive environments where clarity and compatibility accelerate time to market.

These advantages are among the many reasons why existing and emerging international standards should be leveraged throughout PDR efforts. Later in this policy paper, mapping of relevant standards to PDR is provided, illustrating how standards can be applied in practice. By developing and endorsing technologies grounded in robust, consensus-based standards, governments and industry can ensure trust, scalability and innovation are complementary rather than working in opposition. For these reasons the necessity and applicability of standards constantly initiates research into the positive and negative impacts on innovation.[14]

[14] https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100466.pdf

Example: Cross-border adoption in healthcare

The global nature of healthcare makes it a prime example of standards' utility. For instance, ISO 14971:2019, Medical devices – Application of risk management to medical devices, has been adopted as:

- ANSI/AAMI/ISO 14971 in the United States,
- EN ISO 14971 in Europe, and
- JIS T 14971 in Japan.

This coordinated adoption supports global regulatory alignment and facilitates trade while ensuring patient safety.

Similarly, international standards in multimedia can:

- Guide ethical AI deployment,
- Define provenance and authenticity protocols, and
- Protect public trust through verified digital content.

3.2 AI and multimedia authenticity: Standardization in practice

International standards are particularly critical in addressing five key areas of multimedia authenticity:

1. Content provenance,
2. Trust and authenticity,
3. Watermarking,
4. Asset identifiers, and
5. Rights declaration.

The "Technical Report on AI and Multimedia Authenticity Standards: Mapping the Standardization Landscape" provides a comprehensive overview of the current landscape of standards and specifications related to digital media authenticity and artificial intelligence in five clusters. This policy paper has concentrated on three of the five clusters raised by the aforementioned paper: content provenance, trust and authenticity, and watermarking, because these are the most relevant to the issues raised by misinformation and disinformation.

Notably, two separate mapping exercises – one socio-technical/policy and one technical – produced overlapping results, demonstrating strong cross-domain consensus.

Content provenance

Standard number	Responsible group	Title
ISO 22144	ISO TC 171/SC2	Content Credentials
ISO 21617-1:2025	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust Part 1
	Originator Profile	Originator Profile
	Open Provenance	PROV
	C2PA	Content Credential
	Creation Assertions Working Group, as part of DIF	CAWG Metadata

Trust and authenticity of information

Standard number	Responsible group	Title
As yet unnamed	ITU-TSG13 – ISO/IEC JTC 1/SG 29	H.MMAUTH: Framework for authentication of multimedia content
ISO/IEC TR 24028:2020	ISO/IEC JTC 1/SC 42	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
ITU-T Y.3054	ITU-T	Framework for trust-based media services
JPEG Trust Part 2	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust Part 2
ISO/CD 22144	ISO	Authenticity of information — Content credentials
	Creation Assertions Working Group, as part of DIF	CAWG Metadata
	Open Provenance	PROV

Watermarking

Standard number	Responsible group	Title
ISO/IEC 23078-1:2024	ISO/IEC JTC 1/SC 34	Information technology — Specification of digital rights management (DRM) technology for digital publications Part 1: Overview of copyright protection technologies in use in the publishing industry
SMPTE ST 2112-10:2020	SMPTE	Open Binding of Content Identifiers (OBID)
JPEG Trust Part 3	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust Part 3
2413-PLEN	ITU-T SG17	X.ig-dw: Implementation guidelines for digital watermarking
ISO/IEC TR 21000-11:2004	ISO/IEC JTC 1/SC 29/WG 11	Information technology — Multimedia framework (MPEG-21) — Part 11: Evaluation Tools for Persistent Association Technologies
IEEE P3361	IEEE	IEEE Draft Standard for Evaluation Method of Robustness of Digital Watermarking Implementation in Digital Contents
	NIH	A Review of Medical Image Watermarking Requirements for Teleradiology
TR 104 032	ETSI	Securing Artificial Intelligence (SAI)

Other relevant standards

To build trust in AI-generated multimedia there also needs to be assertion that AI bias has been avoided, risk has been fully considered and management systems meet requirements. Internationally recognized standards play a part here too:

Standard number	Responsible group	Title
ISO 24027:2021	ISO	Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
ISO 42001:2023	ISO	Information technology — Artificial intelligence — Management system
ISO 23894:2024	ISO	Information technology — Artificial intelligence — Guidance on risk management
ISO 12791:2024	ISO	Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks

As noted earlier, this policy paper focuses on three key areas: content provenance, trust and authenticity of information, and watermarking. To gain a more comprehensive understanding of the broader standardization landscape, we recommend that this paper be read in conjunction with the accompanying technical pillar report. The pillar report provides an in-depth analysis of two additional areas – asset identifiers and rights declarations – which are also critical to addressing multimedia authenticity challenges. The paper also includes practical recommendations about how and where these standards can be applied, offering valuable guidance for both policymakers and implementers. https://www.worldstandardscooperation.org/what-we-do/amas/

3.3 Conformity assessment: From standards to assurance

Conformity assessment is the process by which conformance with standards or compliance with regulations is verified through methods such as testing, inspection, certification or auditing. Governments and regulators rely on certification and conformity assessment results to determine whether products comply with established requirements of mandatory national technical regulations or voluntary standards. Underpinned by International Standards, such as the ISO/IEC17000 family, conformity assessment is one of the three core pillars (alongside technical regulations and standards) governed by the WTO TBT Agreement. Whether this relates to a product, service, process, claim system or person(s) the whole process provides independent assurance, improves transparency, bolsters supply chain integrity, enhances efficiency and trade facilitation and conformity verification.

In March 2024, the WTO TBT Committee published non-prescriptive practical guidelines to support regulators in the choice and design of appropriate and proportionate conformity assessment procedures with the aim of bringing about convergence.. The underpinning principles are that they be:

- **Non-prescriptive** – they are voluntary and non-binding on WTO members,
- **Neutral** – they allow for different approaches to conformity assessment procedures by regulators across throughout WTO membership,
- **Flexible** – they are intended to allow for innovation in approaches and tools in the field of conformity assessments, and
- **Complementary** – they contribute to the ongoing work of governments, regulators, accreditation bodies, and others at national, regional and international levels, rather than replace existing work and guidance.

With regards to international standards, the guidelines state: "Pursuant to Article 5.4 of the TBT Agreement, Members shall use relevant guides or recommendations issued by international standardizing bodies. For example, Members may make use of conformity assessment standards, such as the ISO Committee on Conformity Assessment (CASCO) toolbox.

Nevertheless, regulators are not limited in their choice of international standards, guides, or recommendations for conformity assessment."

We suggest that regulators consider the TBT Committee's recommendations and the guiding principles when developing conformity assessment schemes for emerging domains, such as multimedia content authentication.

Example: EU AI Act

An area where a similar approach is being adopted is with the EU Conformity Assessments under the proposed EU Artificial Intelligence Act (EU AI Act). Conformity assessments (CAs) are a central mechanism to ensure that high-risk AI systems comply with the regulation's requirements before they are placed on the EU market or put into service. It covers the following areas:

- Risk management system,
- Data governance,
- Technical documentation,
- Record keeping,
- Transparency and provision of information,
- Human oversight, and
- Accuracy, robustness and cybersecurity.

This framework raises the question: should a similar approach be developed for generative AI and multimedia content authentication?

As regulators and legislators design governance mechanisms in this space, they will need to assess which types of conformity assessment provide the most appropriate and effective means of promoting trust, accountability and interoperability, while preserving space for innovation.

Considering both the WTO TBT Committee's guidance and the EU model offers a strong foundation for developing robust conformity assessment schemes to tackle challenges such as misinformation, disinformation, deepfakes and the authentication of multimedia content, without stifling technological advancement.

3.4 Summary

To manage the complex risks associated with multimedia authenticity, misinformation and GenAI, there is a need to adopt a coordinated, standards-based approach. International standards offer a trusted, proven and globally accepted framework to guide regulatory development, support compliance and foster innovation.

When combined with robust conformity assessment mechanisms, these standards:

- Promote mutual recognition throughout jurisdictions,
- Enable interoperability and trust,
- Reduce duplication and resource inefficiency, and
- Protect consumers and uphold public policy objectives.

Ultimately, the PDR framework outlined earlier is only as effective as the standards and assurance systems that support it. Later in this paper, we explore how these tools can be applied practically throughout various domains and stakeholder groups to ensure AI-driven multimedia content remains authentic, ethical and trustworthy.

Section **04**

TECHNOLOGICAL SOLUTIONS AND GUIDANCE

4.1 The role of content provenance in combatting misinformation

Technological solutions that ensure content provenance are fundamental to verifying the authenticity of multimedia. These tools aim to enable the ability to record information about the origin, history and transformation of media over time, creating a transparent digital trail that can help prevent the viral spread of misinformation and rebuild public trust.

The Coalition for Content Provenance and Authenticity (C2PA) is a coalition of technology companies and media organizations with a mission to develop open technical standards for digital content provenance known as Content Credentials. Comprising more than 300 members, the coalition is headed by a steering committee consisting of Adobe, Amazon, BBC, Google, Intel, Meta, Microsoft, OpenAI, Publicis Groupe, Sony and Truepic. Both the EU's 2022 Strengthened Code of Practice on Disinformation and the Partnership on AI's framework for Responsible Practice for Synthetic Media has identified the project as a possible way to increase transparency and authenticity in digital content.

Another leading example in this space is the Content Authenticity Initiative (CAI), which tackles technical, policy and educational challenges in provenance through its promotion of Content Credentials. The CAI comprises a wide alliance of technology companies, academic institutions, media organizations and NGOs, working together to promote adoption of provenance standards globally.

The field of provenance standards is still maturing, but collaborative initiatives by CAI, C2PA, ISO, ITU and IEC are advancing rapidly. They include tools and methodologies for tracking content origins, detecting alterations and establishing trust in digital media.

Content Credentials is being accelerated to become an ISO standard; ISO/CD 22144, Authenticity of information – Content credentials, and as a result, it could soon be officially recognized as a global standard for content provenance and authentication. It provides for cryptographically signed metadata describing the provenance of media that can be attached to the media content during export from software or even at creation time on hardware. With the use of Durable Content Credentials two additional layers of preservation for the retrieval of Content Credentials can be incorporated by adding a digital watermark to the media and implementing a robust media fingerprint matching system.

4.2 Complementary initiatives

WITNESS

For more than 30 years, WITNESS has worked to empower people to use video and technology in the defense of human rights and share trustworthy information. It has recently raised concerns over the risks posed by AI-generated media, particularly the creation of hyper-realistic simulations that can mislead audiences.

WITNESS's focus is not solely on standards, but the organization supports the adoption of frameworks like CAI and C2PA to guide the ethical use of watermarking, labelling and verification systems, which helps balance authenticity with human rights and accessibility considerations.

MAVEN

The MAVEN consortium aimed to integrate content authentication and multimedia analysis tools into a unified platform focused on 'search and verify' functions. The initiative has, however, seen limited uptake, possibly due to competition with better-publicized alternatives, despite its strong foundational objectives.

JPEG Trust

JPEG initiated development of a new International Standard, ISO/IEC 21617-1:2025, Information technology — JPEG Trust. Presented in three parts, it specifies a framework for establishing trust in media that includes aspects of provenance, authenticity, integrity, copyright, and identification of assets and stakeholders.

IEEE Global Initiative on AI Ethics

The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems emphasizes four pillars: global orientation, interdisciplinary collaboration, inclusivity, and practical ethics. This initiative promotes standards, toolkits and certification tools and encourages adoption of the IEEE 7000 Series.

Grassroots and human rights initiatives

Organizations such as the Guardian Project and OpenArchive are leveraging mobile apps like ObscuraCam, InformaCam and ProofMode to support cryptographically verifiable photo, video and audio capture, which enhance documentation for journalism, activism and archiving.

4.3 Emerging commonalities

Throughout these varied initiatives, common features are emerging, including digital signatures, provenance tracking mechanisms, and standardized metadata models. These elements are increasingly seen as essential for operationalizing and regulating multimedia authenticity, especially with the growing use of Public Key Infrastructure (PKI). As streaming platforms, content creators and social media services seek to combat fraud and ensure trust, integration of these features is becoming vital.

Section **05**

SUPPORTING REGULATORY DEVELOPMENT AND CONFORMANCE: CHECKLISTS FOR POLICYMAKERS AND TECHNOLOGY PROVIDERS

To build trust in multimedia authenticity, the following checklist is provided for use by regulators and technology providers when designing regulations and enforcement frameworks or developing technological solutions. It can help align expectations, identify gaps, promote responsible innovation and enable conformity.

Area	Questions for regulators	Questions for technology providers
Scope	What content types are covered?	What types of content do you provide?
	Which ministries or agencies need to be involved?	Are your tools tailored to meet sector-specific regulations?
Regulatory requirements and enforcement	Will the measures be voluntary or mandatory? What are the penalties?	Are there standards or/and conformity assessment schemes that you must comply with? Are you prepared to meet them?
	What is your enforcement capacity? Is there a regulatory body/bodies?	Are you aware of the relevant regulatory authorities?
Standards support	Which voluntary international standards can reinforce your approach or help you achieve your objectives?	Are there relevant standards or conformity assessments to support safe and secure development?
	How can the QI system and relevant institutions help to achieve your objectives?	How can the QI system enhance your solution's credibility?
Technological options	What tools are available for different stages of the content lifecycle? Are they underpinned by standards?	Are your tools evolving with legislative and technical developments?
	What are their benefits and limitations?	Do you clearly communicate the strengths and limitations of your tools?

Below are some additional checklists that can be used by regulators, policymakers and technology providers in situations such as election campaigns, natural disasters and crisis management, with the order in which they should be used.

Action	Document	Description	Use for regulators, legislators and policymakers	Use for technology providers and implementers
One: Begin with this checklist to get an overarching view.	Initial checklist.	Should be used as a starting point.	This checklist should be used to ensure all relevant stakeholders have an opportunity to provide input about their needs.	It is used to give transparency to what governments, legislators or regulators expect them to be able to answer.
Two: Prepare a PDR to identify key risks. Use it based on the scenario that is emerging. For instance, if it is an election campaign create a PDR for that. If it is something like a natural disaster create a new PDR specific to those risks.	Misinformation, disinformation social media PDR.	A PDR to detail the three pillars.	Use as an aid to protect, detect and respond to the risk of misinformation, disinformation from social media. Can be used to ensure any information from regulators, legislators or during government campaigns is protected to ensure ongoing credibility of information received by the public.	Use as an aid to protect, detect and respond to the risk of misinformation, disinformation from social media that affect companies and solution providers.
Three: Depending on the output of the PDR a view will have emerged on what the greatest risks are. Use the matrix to select standards to be followed that give the level of assurance or confidence needed.	MMCA Matrix.	This is a colour-coded matrix, which lists standards, guidance and regulations that exist that can provide different levels of assurance on different topics when different combinations are used. The greater the set that is incorporated the higher the level of assurance.	Can be used by regulators and policymakers who need a starting point to consider those techniques and standards that are available and emerging, which could be referenced or incorporated into a conformity assessment scheme.	Beneficial for organizations wishing to consider self-regulation by the use of techniques outlined, and to consider what level of assurance they may be building.

Action	Document	Description	Use for regulators, legislators and policymakers	Use for technology providers and implementers
Four: In any scenario use this checklist to ensure the correct questions are being asked and checks are being carried out.	Multimedia content authentication checklist.	A spreadsheet listing questions useful for determining the authenticity of multimedia content in a variety of uses.	Can be used by regulators and government departments to verify content. Could be used by policymakers and legislators to encourage auditors and conformity assessment bodies to check what technology solution providers are checking.	Useful for organizations such as news agencies or other media platforms to verify authenticity of content they are sent or consume.
Five: Can be used in parallel with four above or used instead of it if time constraints means that quick answers are needed.	General checklist.	A shorter checklist of questions to consider, and includes a more specific watermarking solutions checklist.	Can be used by regulators, legislators and governments when considering how to assess the authenticity of digital content.	Useful for organizations and media platforms who wish to carry out a quick authenticity check.
Six: Used at any time for any party needing to have specific answers to questions on watermarking.	Watermarking checklist.	A specific checklist that looks at watermarking in more detail.	Can be used by regulators and policymakers, who need a starting point to consider what techniques and standards are available and emerging that could be referenced or incorporated into a conformity assessment scheme for solutions that have a high dependency on watermarking. Could be used by policymakers and legislators to encourage auditors and conformity assessment bodies to check what technology solution providers are checking.	Can be used by technology solution providers developing watermarking solutions to consider what techniques and standards are available and emerging that could be followed to ensure a high-quality product and minimize risk. Useful for any organization wishing to use watermarking or that has a high dependency on watermarking solutions.

Action	Document	Description	Use for regulators, legislators and policymakers	Use for technology providers and implementers
Seven: Use in parallel with any of the other checklists to support decisions about tools available or at a stage where assurance is needed.	Current and emerging techniques for multimedia content authentication guidance.	A supporting document to this paper giving a more technical overview of techniques that are currently being used for authentication.	Can be used by regulators and policymakers, auditors and standards bodies who need a starting point to consider what techniques and standards are available and emerging that could be referenced or incorporated into a conformity assessment scheme.	A useful way for solution providers and organizations wishing to consider self-regulation by the use of techniques outlined, and build confidence by way of assurance techniques.

Section **06**

RECOMMENDATIONS

The following recommendations are intended for international and national policymakers, regulators, the media and technology sectors, and Standards Development Organizations (SDOs). Each can be operationalized swiftly to strengthen multimedia content authenticity and build global trust.

For policymakers and regulators:

- Consider the checklists provided in section four.
- Participate in and collaborate on standard setting and alignment initiatives, especially through multilateral forums to help promote regulatory alignment.
- Consider international standards when developing and implementing regulatory sandboxes to test new technologies, policy approaches and compliance models in controlled environments.
- Adopt a PDR framework based on internationally recognized standards to structure responses to content authenticity challenges.
- Consider data privacy and bias regulations to ensure AI-generated content respects user rights and avoids discriminatory outcomes.
- Support and encourage the development of conformity assessment frameworks specifically targeting multimedia content, incorporating requirements related to AI risks, misinformation, disinformation and deepfakes.
- Consider a conformity assessment and/or certification scheme for multimedia content authentication based on international standards that can give assurance, including relevant testing.

For technology developers and providers, policymakers could request that they consider the following:

- Adopt a PDR framework based on internationally recognized standards to structure responses to content authenticity challenges.
- Align with and monitor international standards and best practices to meet regulatory requirements and future-proof innovation pipelines.
- Assign a standards liaison or champion within your organization to track updates, ensure compliance and guide integration of emerging requirements.
- Consider the integration of strong cryptographic protocols, such as PKI, to enable secure multimedia authentication and content integrity.
- Leverage secure timestamping, tamper-evident hashes and digital signatures to verify content authenticity while preserving user privacy.

Section **07**

CONCLUSION

This policy paper has explored the pressing challenges that AI-generated content and multimedia manipulation pose, particularly in the context of misinformation, disinformation and deepfakes. It underscores the urgent need for coordinated global action supported by robust international standards and conformity assessment frameworks.

By focusing on three key areas – watermarking, content provenance and authenticity – and by leveraging tools such as the PDR framework, this paper outlines actionable steps for regulators, industry and standards bodies to collaboratively address the risks while preserving innovation.

Importantly, the recommendations and supporting checklists provided aim to bridge the gap between policy and practice, enabling generative AI and related technologies to be used safely, ethically and inclusively. When implemented cohesively throughout developed and developing contexts, these measures can help ensure that multimedia content remains trustworthy, verifiable and aligned with public interest.

In conclusion, the effective and harmonized use of international standards, supported by practical guidance and certification, offers a credible path towards a secure, transparent and innovation-friendly digital information ecosystem.

ANNEX 1

The types of misinformation, disinformation and malinformation are extensive. They include areas such as:

Fabricated content	Usually, 100 % false and designed to deceive and do harm.[15] Distinguishing between the real and fabricated content is extremely difficult. Exposure to sophisticated deepfakes used to promote fabricated content can deeply impact trust in the messages citizens receive.
Manipulated content	Genuine information or imagery that has been distorted. These types of content often manipulate genuine content by doctoring an image, or use sensational headlines or click bait.
Imposter content	Impersonation of genuine sources, very often using the branding of an established agency or a reputable news agency. This form of disinformation takes advantage of the trust people have in a specific organization, a brand or even in a person. Adversaries will use phishing and smishing messages using a well-known brand in an attempt to create an impression that the recipient(s) are receiving legitimate content.
Misleading content	Misleading information is created by reframing stories in headlines. This typically uses fragments of quotes to support a wider point, often citing statistics in a way that aligns with a position. Alternatively, it can be the deliberate decision not to cover something because it undermines an argument. When making a point, everyone is prone to drawing out content that supports their overall argument.
False context	Factually accurate content combined with false contextual information, such as the headline of an article failing to reflect the content. Basically, the genuine content has been reframed. False context images are a low-tech but still a powerful form of misinformation and disinformation.
Satire and parody	Humorous but false stores passed off as true; there is no intention to harm, but readers may be fooled. What was once treated as a form of art, is now vigorously used to intentionally spread rumours and conspiracies. It is difficult to police as the perpetrators argue they are merely doing something that shouldn't be treated seriously or literally. The danger of this type of misinformation and disinformation is in the method and speed with which it gets re-shared. In doing so it is often reshaped or reframed and a wider audience loses the connection with the original messenger, failing to understand it as satire.
False connections	Where headlines, visuals or captions, such as sensationalist and click bait headlines don't support the content of an article. At face value this type of content could be perceived as merely irritating, but when efficiently practiced, it has the ability to undermine trust in the media and to promote polarization. As the need to direct traffic to sites grows, it is likely that the relationship between trust and news agencies will diminish.
Sponsored content	Advertising or PR disguised as editorial content. This may appear to be a low impact use of misinformation and disinformation but carries the potential for conflict of interest for genuine news organizations. When consumers are unable to readily identify content as advertising, it can be argued that they are being deliberately mislead through poor labelling.

[15] This type uses false content such as the example of a deepfake audio clip of London mayor Sadiq Khan that was widely circulated on social media in November 2023. The actors used a simulation of the mayor's voice allegedly calling for pro-Palestinian marches to take precedence over Remembrance weekend commemorations on the same day.

Propaganda	Content used to manage attitudes, values and knowledge. Propaganda has always been used as a systematic attempt to shape perceptions, manipulate cognitions and direct behaviour to achieve a response that furthers the desired intent of the propagandist. Traditionally propaganda has involved a complex set of messages each building on the other. Now propaganda uses AI, bots, trolls and fake news sites to disseminate its messages widely and quickly. As a method its effect is more direct and immediate.
Error	A mistake made by established news agencies in their reporting. Errors have existed in news for as long as news has existed. The problem that misinformation and disinformation poses for news agencies is again related to speed. The effort to be the first to present a breaking story minimizes the time for authenticity checks. News agencies are then at the mercy of AI-generated or deepfake content sent from an allegedly legitimate reporter.

ANNEX 2

The nature of the problem impacts many stakeholders, including:

Voters: The intentional dissemination of AI-enhanced misinformation promulgated without any multimedia authenticity during elections increasingly affects voters. This usage serves to deliberately confuse voters and create bias leading to skewed election results in democracies. More widely, such actions undermine public confidence in authority organizations and conventional media, leading to suspicion and disillusionment.

Consumers (consistently impact): When AI tools like predictive analytics and automated advertising targeting are used in consumer scenarios it can have benefits for the consumer and the company. The tools can open up unprecedented efficiency and customer insights, and provide personalized customer experiences. Unfortunately, this also gives rise to negative effects. Consumers can suffer from AI fatigue, whereby the barrage of AI-powered content leads to feelings of inauthenticity and a longing for genuine human connection. This is magnified when content has not been authenticated and results in the consumer becoming a victim of fraud.

Individuals can suffer financial loss or personal harm when malicious actors use unauthenticated multimedia to create fake content for scams or for manipulation purposes. Furthermore, unauthenticated content can be used to track users, steal personal information or spread malware. The use of AI-generated pop-ups that are tracking the shopping patterns of individuals are, by their nature, a coercive force intended to create the urgency to purchase. When pop-ups are maliciously attacked, they can produce instantly threatening messages. Consumers can also be misdirected to sites that produce multimedia content purportedly from genuine advocates of a product.

Adversaries have used AI to generate images that look like celebrities or create audio clips that mimic their voices with such efficiency they are indistinguishable from the genuine article. This often affects the most vulnerable in society who, for example, may for reasons associated with mental health conditions, seek products for quick weight loss or to alleviate anxiety and depression. Similarly, misinformation and disinformation using scientific-sounding articles or videos by so-called medical experts in the field of cancer treatment have for a long time been rife on digital platforms. Claims made that a herb or some alternative therapy either replaces the need for chemotherapy or can alleviate symptoms are common and offer false hope. When these videos incorporate a deepfake of a known authority figure or celebrity purportedly endorsing the product then the persuasive effect increases.

Investment scams are on the rise, and include a recent Facebook example that used a deepfake of respected British financial adviser, Martin Lewis, along with tech billionaire, Elon Musk, promoting a non-existent bitcoin investment scheme. A second involved ITV political analyst and commentator, Robert Peston, also seen recommending a cryptocurrency investment opportunity.

Conceptually this type of misinformation and disinformation ungoverned by any level of multimedia authenticity is predicated on manipulation of human emotions. Consumers who are more likely to fall prey to this are seduced by the idea that a brand is endorsed by an authority figure or celebrity with similar values to their own.

Responding to this issue, Facebook and Instagram owner Meta is set to introduce facial recognition technology to try to crack down on scammers who fraudulently use celebrities in adverts.

Politicians (consistently): For many of us authenticity, when it comes to politicians, is a cornerstone in our evaluations of political candidates and our voting decisions. Our determinations are based on how much we view TV news, the political accounts read or viewed on social media, and candidate profiles. Most people will have their own political attitudes and ideas, but much also depends on specific impressions we derive from the media. That in turn, informs our perceptions of politicians as more or less authentic than their opponents.

Few people have the opportunity to have direct conversations with politicians. As a result, evaluations of a politician's authenticity, trustworthiness and integrity are dependent on impressions formed by media information. In the early days of television interviews with politicians individuals felt empowered to make their evaluation of the politician through the perception that they personally knew the personalities on the screen. Today, social media and populism have enhanced what can be described as a mutually enforcing relationship because of the direct and immediate communicative style. A candidate's self-presentation on social media is a powerful tool, which politicians can use to give the illusion of speaking directly to citizens in a more personal way without the limitations of traditional and institutionalized media.

This should be a positive transformation until we consider the risk of a lack of multimedia authenticity or the concern that the spread of fake news on digital platforms undermines the quality of democratic governance. It is a factor that can be used by politicians, for and against them.

Artists (financially): Another important concern is the large-scale dissemination of AI-authored content in the artworld, exacerbating the already significant problem of digital misinformation. AI tools offer scammers, con artists and criminals a powerful and effective way to create artificial content or false information, including articles, voices, images, photos, videos, songs and artworks, etc. When artificially created in the likeness or the style of the original creators it can be difficult to detect as fake or false. Besides the deliberate misuse of AI tools for nefarious purposes by such actors, authenticity rapidly diminishes as AI-authored content can be produced much faster than purely human-authored content.

Everyday citizens: The outbreak of the COVID-19 pandemic prompted a wave of fake news stories. Misinformation and disinformation proliferated globally with erroneous advice on how to treat the virus putting lives at risk. Whether this was President Trump telling a press conference that the idea of injecting COVID-19 patients with disinfectant "sounds interesting to me" and that "then I see the disinfectant where it knocks it out in a minute. One minute!," or claims that 5G masts were somehow linked to COVID-19 were widely reported at the time. This resulted in at best confusion and at worst mistrust of the authorities attempting to control the situation. In a time of panic and isolation, citizens were highly susceptible to such stories, despite many commentators refuting bogus claims. The sharing of misinformation affected people's psychological well-being and also potentially their wider health.

Social media played a significant role in how individuals perceived the safety of vaccines, with fake stories ranging from claims of harmful ingredients to conspiracy theories that governments used the vaccines to control populations. The effect of unjustifiably influencing a person's decision-making can have consequences that are ultimately catastrophic.

The ease and rate with which individuals and groups with differing agendas used social media to spread misinformation and disinformation led the World Health Organization to coin the phrase "infodemic" while others used the phrase "disinfodemic". Myth-busting campaigns became necessary, especially to combat disinformation that at its core had racist or xenophobic undertones, such as suggestions that people of African descent were immune

Young people: Research undertaken by the UK Safer Internet Centre in 2021 explored how "Half of young people encounter misleading content online on a daily basis". Alongside this, the research also found that "48 % of young people are seeing misleading content every day, with more than one in 10 seeing it more than six times a day – often leaving them feeling annoyed, upset, sad, angry, attacked or scared".[16]

This situation is similar to an addiction where the dependent individual can rationalize the risks and harms they face but cannot break free of the dependency.

[16] https://saferinternet.org.uk/online-issue/misinformation and https://www.getsafeonline.org/personal/news-item/half-of-young-people-encounter-misleading-content-online-daily/

Youngsters, tending to have lower media literacy than adults, are less likely to think critically about news or have sufficient awareness to challenge multimedia authenticity. The dangers they face from intensive exposure to online platforms and the content on offer makes them susceptible to situations that foster anxiety, produces lowered self-esteem, embeds radical opinions (which then pose serious consequences for their beliefs and actions), introduces false memories and can manifest in a catastrophic outlook.

Harmful content is viral and especially dangerous with its interrelationship to other manifestations that social media produces, such as idealization and unrealistic views of other youngster's lives. With a lack of control and governance related to content, misinformation and disinformation exploits the void created by a lack of authenticity controls. Here even simple images are manipulated with filtering producing seemingly realistic portrayals of perfect features and physiques. Any youngster sensitive to body image issues, feeling unable to compete with the flawless images they view and the need to conform or 'measure-up', is even more vulnerable to harmful content promoting self-harm, anorexia, bulimia or suicide-related subject matter.

Cyberbullying and child grooming are ever more proficiently facilitated using emerging technological changes by perpetrators.

ANNEX 3

Deepfakes: Categories and threat vectors

Deepfakes are manipulated or entirely generated synthetic media created using GenAI (e.g. GANs, VAEs, transformers). They are classified by media type and intent.

Type	Method	Threat
Audio deepfakes	Voice cloning: Mimicking an individual's voice using a small sample (e.g. impersonating a CEO). Synthetic speech generation: Creating fake speeches or conversations.	Social engineering (CEO fraud), misinformation, phone scams.
Visual deepfakes	Face swapping: Replacing one person's face with another in video or image. Lip syncing: Altering lip movements to match new audio. Facial expression manipulation: Changing emotions or actions.	Disinformation campaigns, reputation damage, blackmail.
Video deepfakes	Full body reanimation: Entirely generating body gestures and movements. Pose transfer: Mapping one person's pose onto another's body.	Threats: Political manipulation, false confessions, espionage.
Textual deepfakes	Synthetic news/blogs: Generated fake articles or documentation. Fake chatbots/emails: Impersonation in text-based conversations (e.g. phishing).	Fake news propagation, automated trolling, phishing.
Image deepfakes	AI-generated personas: Non-existent faces used in scams or surveillance evasion. Image-to-image translation: Altering visual style/content of images (e.g. removing objects, changing backgrounds).	Sockpuppetry, fraud, fake IDs, misinformation.

Cyber-attacks powered by generative AI

GenAI enables new vectors for traditional and novel cyber-attacks.

Type	Method	Threat
Phishing and social engineering	Spear phishing at scale: AI-generated, customized phishing emails. Voice phishing (vishing): Cloned voice used to deceive targets. Deepfake video phishing: Fake Zoom/ Teams calls mimicking executives.	Credential theft, unauthorized access, BEC (Business Email Compromise).
Malware and exploit generation	Code generation for malware: AI generates polymorphic malware or shellcode. Obfuscation and evasion: GPT-like models create undetectable variants of known malware.	Endpoint compromise, data exfiltration.
Misinformation and disinformation attacks	AI-generated fake news: Large-scale narrative manipulation. Synthetic influencers: Bots with synthetic personas spreading propaganda.	Election interference, economic manipulation, reputational harm.
Impersonation and identity fraud	Synthetic identity creation: Use of GANs to generate fake IDs or entire identity portfolios. Voice/face ID spoofing: Bypassing biometric systems with synthetic inputs.	Bank fraud, KYC circumvention, surveillance evasion.
Data poisoning and model attacks	Training data manipulation: Inserting malicious data into AI model training. Prompt injection attacks: Exploiting LLMs through crafted inputs.	Model degradation, misclassification, unauthorized behaviours.
Content flooding and DDoS of trust	Information overload: GenAI floods platforms with fake content (e.g. reviews, complaints, news).	Overwhelming moderation systems, eroding credibility of information sources.

Hybrid and emerging threat classes

Multimodal deepfakes:

Combining audio, video and text for more convincing deceptions.

Autonomous AI attack agents:

LLMs used to autonomously plan and execute cyber campaigns.

Adversarial example generation:

Images/videos slightly altered to fool AI detection/classification systems.

Synthetic media for sextortion or revenge porn:

Fake intimate imagery used for blackmail.

Defensive considerations

The following are just a few defensive approaches that can help. These are covered in more detail in the checklists.

Detection tools:

Watermarking, fingerprinting, adversarial detectors, forensic tools.

Verification protocols:

Cryptographic signatures, multi-factor verification.

Policy and governance:

AI auditing, legal frameworks, ethical standards.

MISINFORMATION DISINFORMATION SOCIAL MEDIA PDR

Type	Intention	Prevention All	Protection Government specific	Detect All	Respond
Cultivate Fake or Misleading Personas and Websites	Intended to spread disinformation by creating networks of fake personas and websites to increase the believability of their message with their target audience. Typocally fake academic or professional experts, journalists, think tanks, and/ or academic institutions. Fake expert networks use inauthentic credentials to make their content more believable.	Make sure to direct audiences to official websites and trusted sources of information. Make sure your website conveys clear, concise, and current information that people can turn to as a trusted source. Keep online information up to date. Validate all social media accounts for the organization, key representatives, and spokespeople. Verify the sources of articles, papers, and other resources before sharing them.	Government organizations should transition websites to the .gov top-level domain to communicate to the public that the website is genuine and secure using .gov domains that are only available to government departments.	Scan regularly using semantic checkers.	Enforce removal of content using any jurisdiction law or regulation if available.
Synthetic Media and Deepfakes creation	Adversary uses this to convincingly depict someone doing something they haven't done or saying something they haven't said. To use synthetic media technology maliciously as part of a disinformation campaign to share false information or manipulate audiences.	Run awareness campaigns to educate all on how their personal information could be used to generate synthetic media content.Enforce good cyber hygiene practices across both personal and professional accounts. Incorporate publicly available tools, like reverse image search, to verify the source of media content. Add disclaimers to content you share or create that includes synthetic media, even benign uses, to raise public awareness. Develop an incident response plan to deal with deepfake videos or audio clips.		Quickly identify any synthetic media impacting your organization or your message and debunk on official channels, offering evidence, if possible. Use content authenticity tools applicable to sociual media platforms.	Enforce removal of content using any jurisdiction law or regulation if available. Run information campaign to alert victims to the deepfake. Run awareness programs to help citizens identify deepfakes and synthetic content.

Type	Intention	Prevention All	Protection Government specific	Detect All	Respond
Conspiracy Theories (Devising new or amplifying existing ones)	To leverage conspiracy theories that resonate with a target audience by generating disinformation narratives that align with the consipracy perspective. By repeating certain tropes across Using multiple narratives and repeating certain tropes to increas the target audience's familiarity with the narrative and therefore its believability. To effect radicalisation	Keep your website up-to-date with clear, accurate information. Establish both online and offline channels to share information with your peers and partners and collaborate as an amplifying network for trusted information Run awareness campaigns to educate audiences about how conspiracy theories work and common images or figures of speech they may encounter.	Create and maintain a 'Disinformation' or 'Rumor Control' page to immediately debunk fake news or rumours about your department.	Scan sites regularly for items of synthetic media that impacts your organisation. Collaborate with others to share information about adversaries and threat actors.	
Information Flooding and Astroturfing	Increasing audience belief in a message by constant repetition of the same narrative through astroturfing creating the impression of widespread grassroots support or opposition to a message. It's true origin is typically concealed. Using fake or aurtomated accounts to spam social media posts by flooding or firehosing, so that it silences opposing viewpoints, often using many fake and/or automated accounts.	Create a network of trusted communicators in your area to promote authoritative, accurate information. Use more than one channel to communicate so you have alternate ways to share information if your organization is targeted by an astroturfing or flooding campaign. Encourage discussion, debate, and feedback from your constituents through both online and offline forums.	Use officials to create networks of trusted communicators. Leverage other government media channels to raise awareness and combat disinformation. Use popular non government forums to spread good messages through public information narratives/ads that can be hosted by trusted influencers.	Run authenticity checks. If there is suspicion an account is inauthentic 1) check details such as the account creation date 2) profile picture and bio 3) investigate what other sites or accounts they follow 4) investigate posting activity 5) check whether content is posted by suspected bot or troll accounts	

Type	Intention	Prevention All	Protection Government specific	Detect All	Respond
	Often intended to restrict or stop legitimate debate, such as the discussion of a new policy or initiative, and discourage people from participating in online spaces. Information manipulators use flooding to erode the sensitivity of targets through repetition. Intended to create a sense that nothing is true.				
Manipulation of other platforms / small scale community platforms	Intended to create a sense of community by using smaller platforms with less stringent platform and content moderation policies and those that have fewer controls to detect and remove inauthentic content. Using alternative platforms with the intention of capatalising on the less visibility there is on private channels or groups especially those promoting violence. Active intention to recruit followers before going large scale or viral.	Develop training programs so staff know how to respond to external questions and feedback with clear, accurate information and empathy. Ensure enough resources for responding to external audiences. Develop community guidelines and expectations for behavior on social media channels and communicate these to your followers. Create collaborations with partners who have a presence across different communication channels to enable rapid information sharing and amplification.	Encourage questions, feedback, and dialogue from your followers and constituents across communication channels. Develop community guidelines and expectations for behavior on social media channels and communicate these to your staff and followers. Publicise what laws apply in your jurisdiction so the public are aware of the consequences of engaging on these channels if using illegal means.	Run platform checks. 1) check details such as the account creation date 2) profile picture and bio 3) investigate what other sites or accounts they follow 4) investigate posting activity 5) check whether content is posted by suspected bot or troll accounts	Publicise what laws apply in your jurisdiction so the public are aware of the consequences of engaging on these channels if using illegal means.

Type	Intention	Prevention All	Protection Government specific	Detect All	Respond
Manipulation of Unsuspecting Actors	Intended to fool or manipulate prominent individuals and organizations to help amplify disinformation narratives by assumed credibility provided by a secondary spreader often unaware that they are repeating a disinformation actors' narrative or that the narrative is intended to manipulate. Using content that appeals to emotions.	Educate your leadership on how their personal and professional social media presence may be targeted to spread disinformation. Encourage followers to verify sources and assess before subscribing or sharing content through social media.	Protect potential audiences against grassroots disinformation campaigns by proactively debunking or "prebunking," by running awareness campaigns.	Run platform checks. 1) check details such as the account creation date 2) profile picture and bio 3) investigate what other sites or accounts they follow 4) investigate posting activity 5) check whether content is posted by suspected bot or troll accounts	Educate officials on how their personal and professional social media presence may be targeted to spread disinformation Use other platforms to promote messages that clarify political and policy issues.

TOOLS THAT SUPPORT C2PA GUIDANCE

There are several tools and libraries support that support the C2PA (Coalition for Content Provenance and Authenticity) standards, especially through the Content Authenticity Initiative (CAI).

Tool	What it does	When to use
C2PA Tool (Command-Line Utility)	A powerful CLI tool for working with C2PA manifests and media assets.	Reading and displaying manifest data Attaching and signing manifests Creating sidecar files Verifying trust chains Ideal for developers and media professionals working with authenticated content.
CAI Open-Source SDK	A suite of libraries and tools for integrating C2PA into applications:	Use the JavaScript SDK for web-based verification and display of content credentials. Use Rust Library for core implementation used by other SDKs. Use Python, Node.js, and C++/C Libraries in prerelease, for backend or desktop applications. Using these enables creation, verification, and display of Content Credentials.
Web Integration Tools	Tools to embed and display C2PA metadata on websites.	When provenance of data needs to be shown to users, typically beneficial to digital artists, newsrooms and platforms.

CURRENT AND EMERGING TECHNIQUES FOR MULTIMEDIA CONTENT AUTHENTICATION GUIDANCE

In the age of deepfakes, misinformation, and digital forgeries of increasing importance are techniques for Multimedia content authentication. These are techniques that cover the process of verifying the integrity, origin, and authenticity of digital media such as images, videos, and audio. They can be used in multiple applications.

They can be used to proving ownership and originality in the arena of Digital Art & Non-Fungible Tokens (NFTs); for journalism where it is essential to verify the authenticity of user-submitted photos or videos; social media for detecting manipulated or fake content and an area of growing importance is making sure digital media used in court has not been altered.

Techniques	What it does	Key standards and initiatives
Digital Watermarking	Embeds hidden information (e.g., copyright, timestamps) directly into the media. Can be fragile (detects tampering) or robust (survives compression, resizing).	**JPEG Trust (ISO/IEC 19566 series)** Developed by the JPEG Committee (ISO/IEC JTC 1/SC 29/WG 1). Focuses on trust and provenance in digital images. Includes support for digital watermarking and metadata to verify authenticity. **C2PA (Coalition for Content Provenance and Authenticity)** A joint initiative by Adobe, Microsoft, BBC, Intel, and others. Defines a standardized framework for provenance metadata and watermarking in digital content. Although not an ISO standard it is already widely adopted and influential. **ISO/IEC 15444 (JPEG 2000)** Includes optional support for digital watermarking in image compression. Used in applications requiring high fidelity and security, such as medical imaging and digital cinema. **ITU & ISO Collaboration on AI Watermarking** The International Telecommunication Union (ITU) and ISO are working together on standards for: • AI-generated content watermarking • Multimedia authenticity • Deepfake detection

Techniques	What it does	Key standards and initiatives
Digital Signatures	Uses cryptographic techniques to sign media files. Any alteration invalidates the signature, ensuring integrity and authenticity.	**Public Key Infrastructure (PKI)** • Well known and trusted technique that uses a private key to sign content and a public key to verify it. Ensures that the content has not been altered and confirms the identity of the signer and uses common algorithms: RSA, ECDSA, EdDSA. **Detached vs. Embedded Signatures** • Detached: Signature is stored separately from the media file (e.g., .sig file). • Embedded: Signature is embedded within the media file (e.g., in EXIF or XMP metadata). **Hash-and-Sign** • A cryptographic hash of the media is generated and then signed. • Efficient and secure, especially for large files. **Timestamping** • Adds a trusted timestamp to the signature to prove when the content was signed. • Useful for legal and archival purposes. **ISO/IEC 9796 & 14888** Standards for digital signature schemes and message recovery. Applicable to multimedia when combined with hashing and metadata. X.509 Certificates • Used in PKI to bind public keys to identities. • Common in secure email, HTTPS, and digital content signing. **W3C Verifiable Credentials** The framework for digitally signed claims about content or identity. Can be used to verify the authenticity of media creators or publishers. **CAdES, XAdES, PAdES** **C2PA (Content Provenance)**

Techniques	What it does	Key standards and initiatives
Hashing	Generates a unique hash value for a file. If the file changes, the hash changes — useful for tamper detection.	**Cryptographic Hash Functions** These are standard, secure hash algorithms used to generate a unique fingerprint of a file. SHA-256 / SHA-3 (Secure Hash Algorithm) Widely used in digital signatures and blockchain. Any change in the media file results in a completely different hash. Standardized by NIST (FIPS 180-4). Perceptual Hashing (pHash, aHash, dHash) Used for content-based identification — tolerant to minor changes like resizing or compression. • pHash (Perceptual Hash) • Captures the essence of an image or video frame. • Similar-looking media will have similar hashes. • Useful for detecting near-duplicates or slight edits. • aHash (Average Hash) • Simplified method based on average pixel values. • Fast but less robust than pHash. • dHash (Difference Hash) • Based on pixel differences. • Good for detecting structural changes. Note: MD5 / SHA-1 is still used in legacy systems of for non-critical integrity checks. However, it is not very secure and has known vulnerabilities so should be avoided if possible Video Hashing Techniques: • Frame-based Hashing: Where perceptual hashing is applied to keyframes. • Motion Hashing: Captures motion vectors and scene changes. • Temporal Hashing: Considers the sequence and timing of frames. Standards and Frameworks: ISO/IEC 15938 (MPEG-7): Multimedia content description interface — includes descriptors for image and video signatures. ISO/IEC 23000-19 (MPEG-21 Media Value Chain Ontology): Supports content identification and authentication. C2PA: Uses cryptographic hashes to bind metadata and content securely.

Techniques	What it does	Key standards and initiatives
Blockchain-Based Authentication	Stores media metadata or hashes on a blockchain to provide a tamper-proof, decentralized record of authenticity.	**Content Hashing on Blockchain SHA-256** Smart Contracts for Rights and Access The use of tools such as Interplanetary File System and + Ethereum or other chains so that media can be stored off-chain while its hash is held on chain reducing storage costs but maintaining integrity. **W3C Verifiable Credentials** A standard for digitally signed claims about content or identity often integrated with blockchain for decentralized verification. Can be used to verify the authenticity of media creators or publishers. **C2PA (Coalition for Content Provenance and Authenticity)** Although not blockchain-native, it can be integrated with blockchain for immutable provenance tracking. **ISO/TC 307 – Blockchain and Distributed Ledger Technologies** The technical committee who are developing global blockchain standards to cover areas like identity, smart contracts, and data integrity, which are relevant to multimedia authentication.
AI-Based Forensics	Uses machine learning to detect signs of manipulation (e.g., deepfakes, splicing). Can analyze inconsistencies in lighting, shadows, compression artifacts, etc.	**Deepfake Detection** Convolutional neural networks (CNNs) and transformers are used to detect synthetic media looking for inconsistencies in facial movements, eye blinking, lighting, and audio-visual sync. Several GAN-resistant models are being developed to counter anti-forensic attacks **Splicing and Tampering Detection** A method used to detect inconsistencies in compression artifacts, lighting, or shadows. The techniques uses multi-scale CNNs and attention mechanisms to localize tampered regions. **OpenMFC (NIST)** NIST-led initiative to standardize multimedia forensic challenges and benchmarks Focuses on deepfake detection, provenance, and anti-forensics. **C2PA (Coalition for Content Provenance and Authenticity)** A framework for embedding and verifying provenance metadata. Metadata and Provenance Analysis AI models cross-reference metadata with visual content to detect anomalies. Often integrated with blockchain or C2PA frameworks for traceability. (see section on blockchain). **Explainable AI (XAI)** Enhances trust by making AI decisions interpretable to forensic analysts and investigators. Useful wherever transparency is critical.

Techniques	What it does	Key standards and initiatives
Metadata Analysis	Examines embedded metadata (EXIF, timestamps, GPS). Can reveal inconsistencies or signs of editing.	**EXIF (Exchangeable Image File Format)** The widely used standard for digital photography and forensics for storing metadata in image files (e.g., camera model, date/time, GPS). XMP (Extensible Metadata Platform) Supports custom schemas and is used in Content Credentials. Developed by Adobe; allows embedding metadata in various file types. **C2PA (Coalition for Content Provenance and Authenticity)** Tracks content origin, editing history, and ownership and embeds cryptographically signed metadata into media. **MPEG-7 (ISO/IEC 15938)** Multimedia content description interface. It defines descriptors for low-level features (colour, texture) and high-level semantics (events, objects). **SWGDE Best Practices** The Scientific Working Group on Digital Evidence provides guidelines for metadata analysis in digital video authentication. **Dublin Core & IPTC** Standardized metadata tagging used in journalism and digital libraries.

Hash Comparison Chart

The following chart is intended to help differentiate different types of hashing methods depending on the priority.

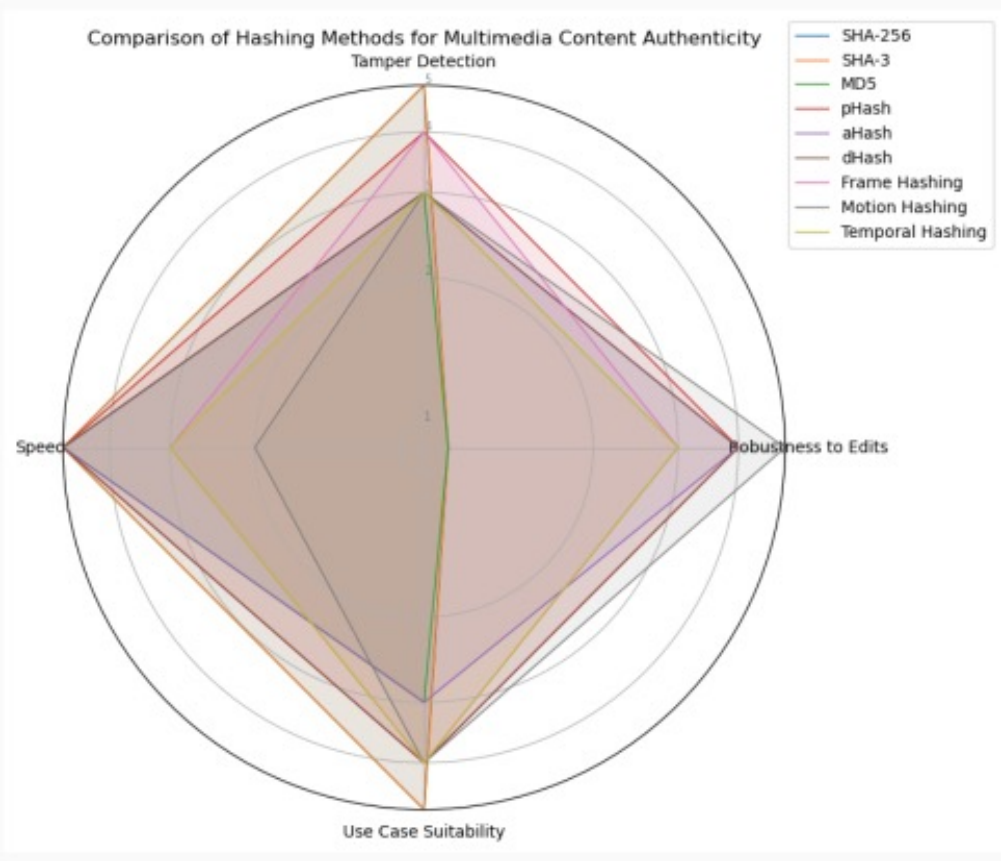
- Robustness to Edits
- Speed
- Tamper Detection
- Use Case Suitability

Interpretation:

Note that each line represents a different hashing method, the further out a method reaches on the axis the better its performance in that category.

Decide what is the most important criteria for the use case in question, for instance, if the main concern is perceptual similarity detection, it is clear that pHash has more to offer. Whereas, SHA-256 is preferable when speed and tamper detection are priorities.

Comparison of Hashing Methods for Multimedia Content Authenticity



GENERAL CHECKLIST

A multimedia content authenticity checklist can help your organisation ensure the integrity and origin of digital content.

This involves verifying the source, history, and any alterations made to the content. You should not miss the checks meant for metadata, source information, and proof of editing or manipulation.

Topic	Check	What to check	Result
Source and Provenance			
1)	Verify the original source.	Can you establish the location, time, and creator of the content.	
2)	Check for metadata.	What embedded information is found such as camera settings, location data, and timestamps.	
3)	Review file details.	Examine file names, versions, and other attributes for clues about the content's history.	
Editing and Manipulation			
1)	Identify potential alterations:	Look for signs of editing, such as retouching, cropping, or digital enhancements.	
2)	Assess the impact of edits:	Consider how the alterations might affect the content's meaning and context.	
3)	Document the history of edits:	Note any modifications made to the content and who made them.	
Verification and Validation			
1)	Use authenticity tools.	Utilize software or services that can verify the source and history of digital content.	
2)	Is there need for expert guidance?	If necessary seek guidance from professionals who specialize in content authenticity or media forensics.	
3)	What standards can be used?	Use Content Authenticity Initiative (CAI) and C2PA for industry standards.	

Topic	Check	What to check	Result
Transparency and Disclosure			
1)	Provide context.	Make sure there is a way to label the content as original or altered, and explain any changes that have been made.	
2)	Attribute correctly.	Make sure there credit is being given to the original creator and any individuals or entities involved in the content's creation or editing.	
3)	Share information openly.	Make sure relevant details about the content's origin and history available to the public.	

WATERMARKING SOLUTIONS CHECKLIST

We present here a checklist to be followed when selecting a watermarking solution.

After this checklist is a table providing names of solutions that have been checked by the authors. That table is informative only and no claims are made as to preference. Users of this checklist should ensure that the solution is credible and appropriate for their use.

Watermarking Solution Checklist		
Pre-selection questions		**Response**
	Does the solution offer the ability to handle a range of types of content, such as images, videos, or audio.	
	Is a demonstration accessible.	
	What level of protection does it provide? This will depend on the specific needs of the content. Consider even if not necessary now does it offer forensic watermarking to provide a higher level of security as threats increase.	
Ease of operation and use	Can the solution be easily integrated into the content creator workflow without the need for specialized technical expertise.	
	Can the solution be easily integrated into the content creator workflow significant changes to the workflow process.	

Watermarking Solution Checklist		
Pre-selection questions		**Response**
Pre-selection questions		Response
Cost	What does the cost cover? Does it require extra 'plug-ins' or 'add-ons' that are chargeable? Are updates free? Are there hidden costs?	
Licence	What are the licencing details?	
After down-selection move to these checks:		
Technical specifications	Do a security review to see how robust the solution is to reverse engineering and forgery.	
	Do an evaluation if detection rates under content modifications.	
	Analyse and verify imperceptibility across content types.	
Comparative testing	Conduct side by side tests with your particular content.	
	Use industry standard metrics to evaluate and measure performance.	
Scalability, integration and interoperability	Perform an evaluation of ease of integration with existing systems.	
	Perform an assessment of the solution's ability to handle current content volume and to scale to handle increasing levels of future content.	

Available solutions

The solutions are presented in alphabetical order to avoid any suggestion of bias or preference.

Solution	Use	
DataPatrol	Provides a variety of solutions more geared to device marking and web marking.	
Digimarc	Provides solutions for a variety of use cases and uses GS1.	
Digital Guardian/Fortra	Provides a range of watermarking solutions.	
Dropsend	Provides dynamic watermarking for sensitive documents.	
Friend MTS	Provides watermarking solutions for live sports and other entertainment industries, including subscriber ID watermarking.	
Google DeepMind	Provides SynthID, a system for watermarking AI-generated content.	
IMATAG	Provides watermarking solutions for the media and publishing industry, including forensic watermarking and monitoring services.	
MediaValet	Provides watermarking solutions for protecting media assets, including generating watermarked renditions of images.	
NAGRA	Provides forensic watermarking solutions for protecting digital media and content.	
NoisyPeak	Provides end-to-end watermarking solution to protect audio and visual content which can be applied to existing content items or include additional transcoding and DRM protection. Forensic protection and content tracking.	
Synamedia	Provides forensic watermarking solutions for media and entertainment, including ContentArmor.	

MMCA Checklist		
Area	**Questions**	**Guidance**
Source Verification:		
Authority & Credibility:		
	Is the author, publisher, or sponsor identified and verifiable or can you confirm the identity of, and contact, the person?	
	Are you familiar with this account?	
	Has their content and reportage been reliable in the past?	
	Can their expertise or credentials be verified?	
	Has the source been cited by other reliable sources?	
	What information do you have trust this source?	
	What biographical information is evident on the account?	
	What are their main narratives/discussion points?	
	Does any biographical information conflict with the type of content?	For instance is the content intended for an older age group but the language used suggests it has been created or manipulated by someone younger. This is often identifiable by use of urban vocabulary.
	Can you establish where the uploader is based? (see account history below)	Location is often an indicator of political motivation and can be detectable when there is a contradiction between location claimed, biographical details and verification of uploader residence.
	Does it link anywhere else?	
Account History (if applicable):		
	How active is the account?	
	How active is the uploader on the account?	
	What type of content has been previously uploaded?	
	Are there any inconsistencies or warning signs in their account history?	

MMCA Checklist		
Area	**Questions**	**Guidance**
Source's Social Network Connections:		
	Who are their friends and followers?	
	Who are they following?	
	Who do they interact with?	
	Are they connected to any known misinformation channels or individuals?	
	Look for other accounts associated with the same name/username on other social networks in order to find more information.	If you find a real name, use can use people search tools to find the person's address, email and telephone number: Pipl.com White Pages Spokeo WebMii Check if a Twitter or Facebook Verified account is actually verified by hovering over the blue check. If the account is verified by Twitter or Facebook, a popup will say "Verified Account" or "Verified Page." Check LinkedIn, to find out about the person's professional background.
Content Examination:		
Accuracy & Consistency:		
	Can the information be verified with other reliable sources?	
	Do a time check.	You can use tools like Wolfram Alpha to perform a search on specifics like the weather that day and then check the weather information on the day and the location where the event allegedly happened. Veryfying the weather conditions from the same from the local weather forecasts is a good check to run. Check to see if any earlier pieces of content from the same event predate what you are looking at. You can use tools that provide timestamps and use video and image search with Google, Tin Eye and YouTube for example. Don't dismiss commonsense checks for images and video, look and listen for anything that confirms or refutes date/ time this could be clocks on a shelf, television screens showing a program that was never shown that day, or a newspaper pages with a date that has yet to occur according to the content under scrutiny.

MMCA Checklist		
Area	**Questions**	**Guidance**
	Do a location check.	You can use tools like Wolfram Alpha to perform a search on specifics like the weather that day and then check the weather information on the day and the location where the event allegedly happened. Verifying the weather conditions shown in the image or video match those reported by tools like Wolfram Alpha.
		You should check if the content includes automated geolocation information?
		Check reference points that you can compare with satellite imagery and geolocated photographs this could be street signs, building signs. Look for anomalies where car registration plates are predominantly registered in a countruy other than the one suggested. Is advertising signage in the correct language for the location? Look for distinctive landscapes that can confirm or refute the geolocation claimed. You could look for sports stdiums, cathedrals and so on. A number of freely available tools can be used such as Google Maps and Google Street View.
	Does the research contain sufficient evidence to back up the claims?	
	Are there any inconsistencies or contradictions within the content itself?	
Signs of Manipulation:		
Images		
	Does the image or video look as though it has been doctored or manipulated?	Use tools to verify the provenance.
	Does the image or video match what the accompanying text says?	Use tools to verify the provenance.
	Are there any obvious alterations or distortions?	Look at things such as lip synching.
Video		
Voice		

MMCA Checklist		
Area	**Questions**	**Guidance**
Search Engine Checks:		
	Perform a reverse image search to see if the image appears in other contexts.	In what contexts does it appear, are any of these inflamatory, prejudicial etc.
	Use search engines to verify the accuracy of claims and information.	
Fact-Checking:		
	Check fact-checking websites to see if the content has already been verified.	
	Submit the content for verification to fact-checkers if necessary.	
Context & Support:		
Cross-Referencing:		
	Check if the content is supported by other reliable sources.	See Accuracy and Consistency section above.
	Verify the information against multiple sources to avoid bias.	See Accuracy and Consistency section above and then check other reliable sources.
Timeliness & Relevance:		
	Is the content relevant to current events or trends?	Look at timestamps.
	Is the content up-to-date and accurate?	Look at timestamps.
Analyse the Author's Perspective and Motivations:		
	Is there any bias or agenda behind the content?	Consider the narrative or tropes used are they common to any particular faction or group.
	What are the potential implications of sharing this content?	Do a risk assessment of the inplications of sharing.

MMCA MATRIX

Essential | Advised

Content Provenance	Strong	Medium	Foundational
	Standard No	Standard No	Standard No
	ISO 22144 Content Credentials	ISO 22144 Content Credentials	ISO 22144 Content Credentials
	ISO 21617-1: 2025	ISO 21617-1: 2025	ISO 21617-1: 2025
	Originator Profile	Originator Profile	Originator Profile
	Open Provenance	Open Provenance	Open Provenance
	C2PA	C2PA	C2PA
	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system
	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management
	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security
	GDPR or other privacy regulation/framework	GDPR or other privacy regulation/ framework	GDPR or other privacy regulation/framework
	OWAS Secure coding practices	OWAS Secure coding practices	OWAS Secure coding practices

Trust and Authenticity of information	Strong	Medium	Foundational
	Standard No	Standard No	Standard No
	ITU-TSG13 – ISO/IEC JTC 1/SG 29 H.MMAUTH: Framework for authentication of multimedia content	ITU-TSG13 – ISO/IEC JTC 1/SG 29 H.MMAUTH: Framework for authentication of multimedia content	ITU-TSG13 – ISO/IEC JTC 1/SG 29 H.MMAUTH: Framework for authentication of multimedia content
	ISO/IEC TR 24028: 2020 Information Technology – Artificial Intelligence – Overview of trust worthiness in artificial intelligence	ISO/IEC TR 24028: 2020 Information Technology – Artificial Intelligence – Overview of trust worthiness in artificial intelligence	ISO/IEC TR 24028: 2020 Information Technology – Artificial Intelligence – Overview of trust worthiness in artificial intelligence
	ITU-T Y.3054 Framework for trust-based media services	ITU-T Y.3054 Framework for trust-based media services	ITU-T Y.3054 Framework for trust-based media services
	ISO/CD 22144 Authenticity of information — Content credentials	ISO/CD 22144 Authenticity of information — Content credentials	ISO/CD 22144 Authenticity of information — Content credentials
	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system
	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management
	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
	Information Sources Authenticity Checklist (ISAC)	Information Sources Authenticity Checklist (ISAC)	Information Sources Authenticity Checklist (ISAC)
	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security
	GDPR or other privacy regulation/framework	GDPR or other privacy regulation/framework	GDPR or other privacy regulation/framework
	OWAS Secure coding practices	OWAS Secure coding practices	OWAS Secure coding practices

Watermarking	Strong	Medium	Foundational
	Standard No	Standard No	Standard No
	ISO/IEC 23078-1:2024 Information technology — Specification of digital rights management (DRM) technology for digital publications	ISO/IEC 23078-1:2024 Information technology — Specification of digital rights management (DRM) technology for digital publications	ISO/IEC 23078-1:2024 Information technology — Specification of digital rights management (DRM) technology for digital publications
	SMPTE ST 2112-10:2020 Open Binding of Content Identifiers (OBID)	SMPTE ST 2112-10:2020 Open Binding of Content Identifiers (OBID)	SMPTE ST 2112-10:2020 Open Binding of Content Identifiers (OBID)
	2413-PLEN X.ig-dw: Implementation guidelines for digital watermarking	2413-PLEN X.ig-dw: Implementation guidelines for digital watermarking	2413-PLEN X.ig-dw: Implementation guidelines for digital watermarking
	ISO/IEC TR 21000-11:2004 Information technology — Multimedia framework (MPEG-21) — Part 11: Evaluation Tools for Persistent Association Technologies	ISO/IEC TR 21000-11:2004 Information technology — Multimedia framework (MPEG-21) — Part 11: Evaluation Tools for Persistent Association Technologies	ISO/IEC TR 21000-11:2004 Information technology — Multimedia framework (MPEG-21) — Part 11: Evaluation Tools for Persistent Association Technologies
	IEEE P3361 IEEE Draft Standard for Evaluation Method of Robustness of Digital Watermarking Implementation in Digital Contents	IEEE P3361 IEEE Draft Standard for Evaluation Method of Robustness of Digital Watermarking Implementation in Digital Contents	IEEE P3361 IEEE Draft Standard for Evaluation Method of Robustness of Digital Watermarking Implementation in Digital Contents
	TR 104 032 Securing Artificial Intelligence (SAI)	TR 104 032 Securing Artificial Intelligence (SAI)	TR 104 032 Securing Artificial Intelligence (SAI)
	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	ISO 24027 : 2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system	ISO 42001 : 2023 Information technology — Artificial intelligence — Management system
	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management	ISO 23894:2024 Information technology — Artificial intelligence — Guidance on risk management

Watermarking	Strong	Medium	Foundational
	Standard No	**Standard No**	**Standard No**
	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks	ISO 12791:2024 Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks
	Legal compliance for jurisdiction	Legal compliance for jurisdiction	Legal compliance for jurisdiction
	Copyright law for jurisdiction	Copyright law for jurisdiction	Copyright law for jurisdiction
	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security	ISO 27001: 2022 Information Security
	GDPR or other privacy regulation/framework	GDPR or other privacy regulation/ framework	GDPR or other privacy regulation/framework
	OWAS Secure coding practices	OWAS Secure coding practices	OWAS Secure coding practices

ISBN 978-2-8399-4721-3