POLICY PAPER: BUILDING TRUST IN MULTIMEDIA AUTHENTICITY THROUGH INTERNATIONAL STANDARDS

政策文書：国際規格を通じたマルチメディアの真正性への信頼構築

CAROL BUTTLE AND CINDY PAROKKIL

キャロル・バトル，シンディ・パロッキル

政策文書
第1版

英和対訳
一般財団法人 日本規格協会

TABLE OF CONTENTS

AI and Multimedia Standards Collaboration 04

Preface 05

Acknowledgement 05

1 **The Context** 06

1.1 Misinformation and disinformation in the age of AI

1.2 Definitions matter: Misinformation, disinformation and malinformation

1.3 Who are the types of perpetrators presenting challenges?

1.4 Deepfakes and cyber-attacks

1.5 AI and multimedia authenticity

2 **The complexities of balancing the regulatory landscape with market needs to build trust in multimedia authenticity** 13

2.1 Why is building trust in multimedia authenticity complex?

2.2 What factors contribute to the difficulties?

2.3 Overview of regulatory landscape

2.4 Bridging the gap between regulation and trust

2.5 Finding practical solutions for governments and industry

3 **The role of international standards and conformity assessment in addressing multimedia authenticity** 23

3.1 The value of international standards

3.2 AI and multimedia authenticity: Standardization in practice

Content provenance

3.3 Conformity assessment: From standards to assurance

3.4 Summary

目次

AIとマルチメディア規格の連携 04

序文 05

謝辞 05

1 背景 06

1.1 AI時代の誤情報と偽情報

1.2 定義の重要性：誤情報，偽情報，悪意の情報

1.3 問題となる加害者の種類は？

1.4 ディープフェイクとサイバー攻撃

1.5 AIとマルチメディアの真正性

2 マルチメディアの真正性への信頼構築における規制環境と市場ニーズのバランスの複雑さ 13

2.1 マルチメディアの真正性への信頼構築はなぜ複雑なのか？

2.2 困難をもたらす要因とは？

2.3 規制環境の概要

2.4 規制と信頼のギャップを埋める

2.5 政府と産業界のための実践的なソリューションの模索

3 マルチメディアの真正性確保における国際規格と適合性評価の役割 23

3.1 国際規格の価値

3.2 AIとマルチメディアの真正性：実践的な標準化 コンテンツの出所

3.3 適合性評価：規格から保証へ

3.4 まとめ

4 **Technological solutions and guidance** 31

4.1 The role of content provenance in combatting misinformation

4.2 Complementary initiatives

4.3 Emerging commonalities

5 **Supporting regulatory development and conformance: Checklists for policymakers and technology providers** 33

6 **Recommendations** 37

7 **Conclusion** 38

Annex 1 39

Annex 2 40

Annex 3 44

4 技術的ソリューションとガイダンス 31
4.1 誤情報対策におけるコンテンツの出所の役割
4.2 補完的な取り組み
4.3 新たな共通点

5 規制策定と適合性支援：政策立案者と技術提供者向けチェックリスト 33

6 推奨事項 37

7 結論 38

附属書1 39

附属書2 40

附属書3 44

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)

AIとマルチメディア規格の連携

AIとマルチメディア真正性規格の連携は，急速に進化するAIによって生成・改変されたメディアの分野における標準化を推進する世界的な取り組みです。ギャップを特定し，新たな規格の開発を推進することで，透明性，プライバシーへの配慮，そして権利尊重を重視した実践を支援します。また，法令遵守を促進し，公共の信頼を守るためのおよび規制フレームワークの構築にも貢献します。

世界標準協力（WSC）[1]が主導するこの連携は，規格開発者，市民社会組織，テクノロジー企業，その他の主要関係者間の対話のための重要なフォーラムとして機能しています。参加組織には，国際電気標準会議（IEC），国際標準化機構（ISO），国際電気通信連合（ITU），コンテンツの出所・真正性に関する連合（C2PA），中国情報通信研究院（CAICT），DataTrails，Deep Media，Witnessなどが含まれます。

この連携は，世界標準化協力（WSC）の後援の下，ITUが招集し，2024年に開催されたAI for Goodグローバルサミットで開始されました。

詳細はこちら（https://aiforgood.itu.int/multimedia-authenticity/）をご覧いただくか，事務局（amas-secretariat@itu.int）までお問い合わせください。

[1] 国際電気標準会議（IEC），国際標準化機構（ISO），国際電気通信連合（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.

序文

本文書は，主に政策立案者および規制当局者を対象としています。合成マルチメディアコンテンツの作成，利用，配信を規制する際の複雑さを，防止，検知，対応を通して分かりやすく解説し，専門知識や専門的理解のレベルが異なる対象者にとって，これらの問題を明確かつ分かりやすく提示することを目的としています。さらに，本文書は，世界的な取り組みを取り上げ，法域をまたいだ規制の一貫性，整合性，および効果的な執行を促進する上で，国際規格が果たす重要な役割と利点を強調することを目的としています。

本文書は，政策立案者や規制当局が規制対象（範囲），規制方法（自主規制か強制規制か），そしてどの程度まで規制するか（努力のレベル）を決定するのに役立つ規制オプションマトリックスを含む，実践的なガイダンスと実行可能な推奨事項を提供しています。また，マルチメディアコンテンツの悪用に起因する誤情報や偽情報の課題への対処に貢献できる，規格，適合性評価メカニズム，支援技術といった様々な支援ツールについても考察しています。同時に，完全合成または部分合成のマルチメディアを社会，政府，そして商業の利益のために積極的かつ合法的に利用できるよう，バランスを取ることの重要性を強調しています。

最後に，本文書には，政策立案者，規制当局，そして技術提供者が活用できる実用的なチェックリストが掲載されています。これらは，規制や執行フレームワークの策定，技術的ソリューションの開発，危機対応戦略の策定に活用できます。これらのチェックリストは，利害関係者の期待を一致させ，重要なギャップを特定し，責任あるイノベーションを支援し，新たな規格やベストプラクティスへの適合を可能にすることを目的としています。

謝辞

本書は，ISOが主導するAMASの政策の柱に基づき，キャロル・バトル氏とシンディ・パロッキル氏が共著で作成しました。本政策論文の作成にあたり，AMASメンバーの皆様からいただいた貴重な貢献に深く感謝申し上げます。なお，誤りはあれば著者自身の責任となります。

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.

第1章

背景

生成型人工知能（GenAI，【JSA注】以下和訳では“生成AI”とする）は，ここ数十年で最も革新的な技術の一つとなる可能性を秘めています。しかし，その可能性を実現するには，生成AIがもたらす計り知れない利点を認識するだけでなく，それに伴う重大なリスクを認識し，管理することも必要です。歴史的に，新興技術の社会への影響は，その導入のスピード，広さ，深さに左右されてきました。生成AIの場合，導入は前例のないペースで加速しており，思慮深い設計と堅牢なガバナンスの重要性が高まっています。

生成AIがあらゆるセクターに統合されるにつれ，政策，規制，規格，コンプライアンス，認証を網羅する包括的なフレームワークの必要性が高まっています。これらのフレームワークは，生成AIシステムの構想段階から設計段階まで，安全策と倫理原則を組み込む必要があります。これは，先進国と発展途上国を問わず，社会と経済を変革する強力な力を持つこの技術の複雑さを政策立案者が理解していく上で，特に分断された世界的な法規制環境において，政策立案者にとって大きな課題となります。国際的な連携の緊急性はかつてないほど高まっています。

> “生成型AIの潜在的な影響とリスクは国境を越え，新たな政策と技術ソリューションをグローバルな視点から必要としています。”
>
> 経済協力開発機構

経済協力開発機構（OECD）は，生成AIはグローバルな視点から理解されるべきであり，それに応じて政策と技術ソリューションが策定されるべきであることを強調しています。18世紀と19世紀の産業革命はイギリスで始まり，その後ヨーロッパやアメリカに広がりましたが，AI革命は世界規模で同時進行しています。各国は今，ガバナンス，医療，防衛，市民社会といった分野における生成AIのメリットを実現すると同時に，リスクを軽減し，市民を悪用から守るという二重の課題に取り組まなければなりません。

ディープフェイク，AI生成テキスト，画像，音声など，人工知能（AI）を用いて生成または操作されたコンテンツである合成メディアは，機会と深刻な課題の両方をもたらします。

誤情報や偽情報の明らかな蔓延に加え，信頼の喪失という問題もあります。合成メディアがよりリアルになるにつれて，本物と偽物の区別が難しくなります。さらに，多くの国で合成メディアの作成と使用に関する明確な法律がないため，法的および倫理的な曖昧さが生じ，同意，所有権，説明責任といった問題が生じます。

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.

デジタルIDの重要性と普及がますます高まる世界では，個人情報の盗難や詐欺のリスクも高まっています。合成IDは，金融詐欺やデジタルIDシステムの操作を目的として作成されることが一般的です。生体認証システムも，顔のディープフェイクや音声クローンによる攻撃を受けやすい傾向があります。

生成AIと合成メディアは様々な可能性をもたらしますが，透明性，情報開示，そして被害軽減に重点を置いた包括的な政策によって支えられて初めて実現可能です。

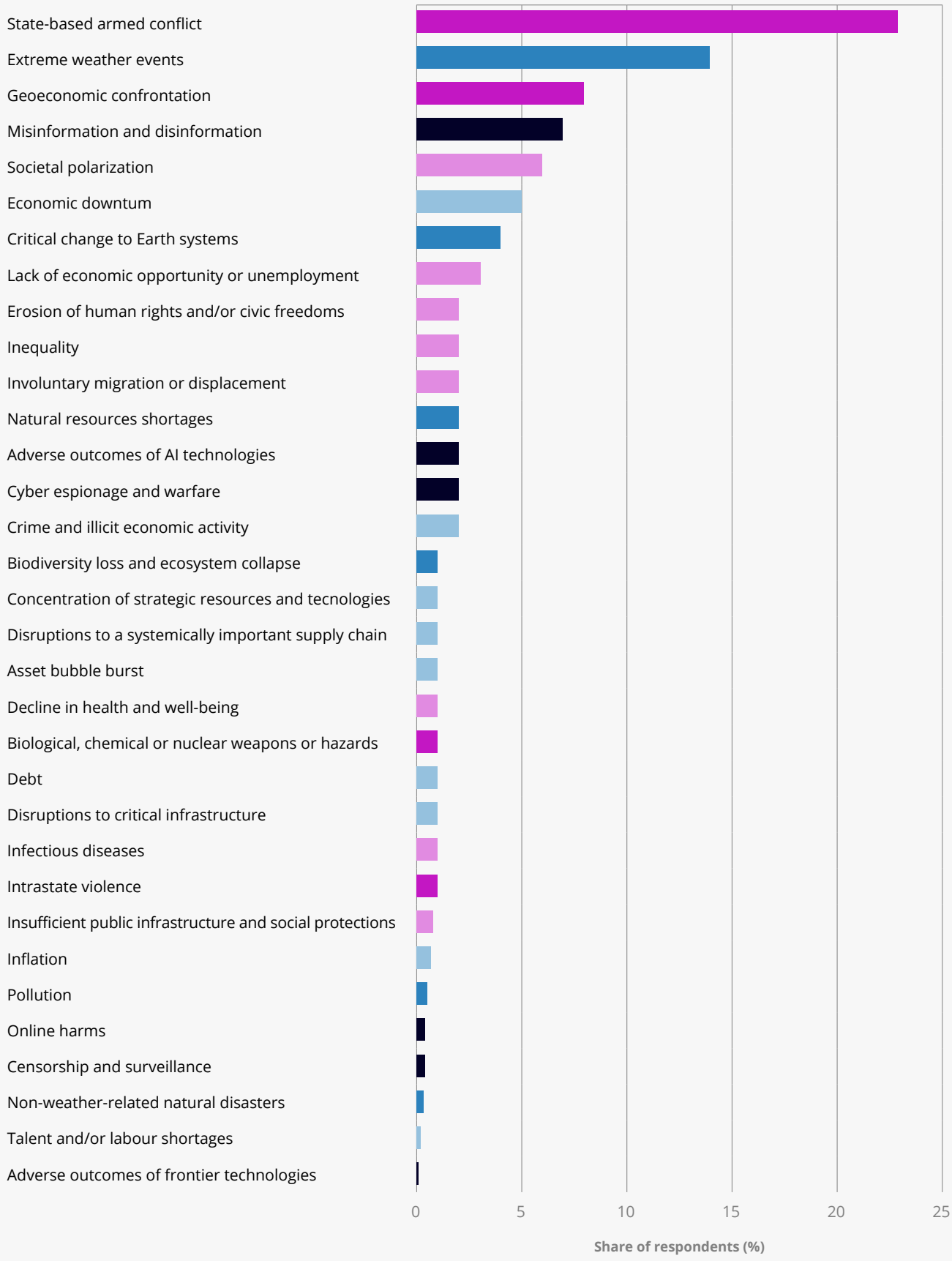
1.1 AI時代の誤情報と偽情報

情報の真正性に関する懸念は，何世紀にもわたって存在してきました。しかし，AIによって加速されたデジタル時代とデジタル環境は，これらの問題を深刻化させ，国境を越えたグローバルな脅威へと変貌させ，国民の信頼，国家安全保障，そして民主主義制度に重大な影響を及ぼしています。

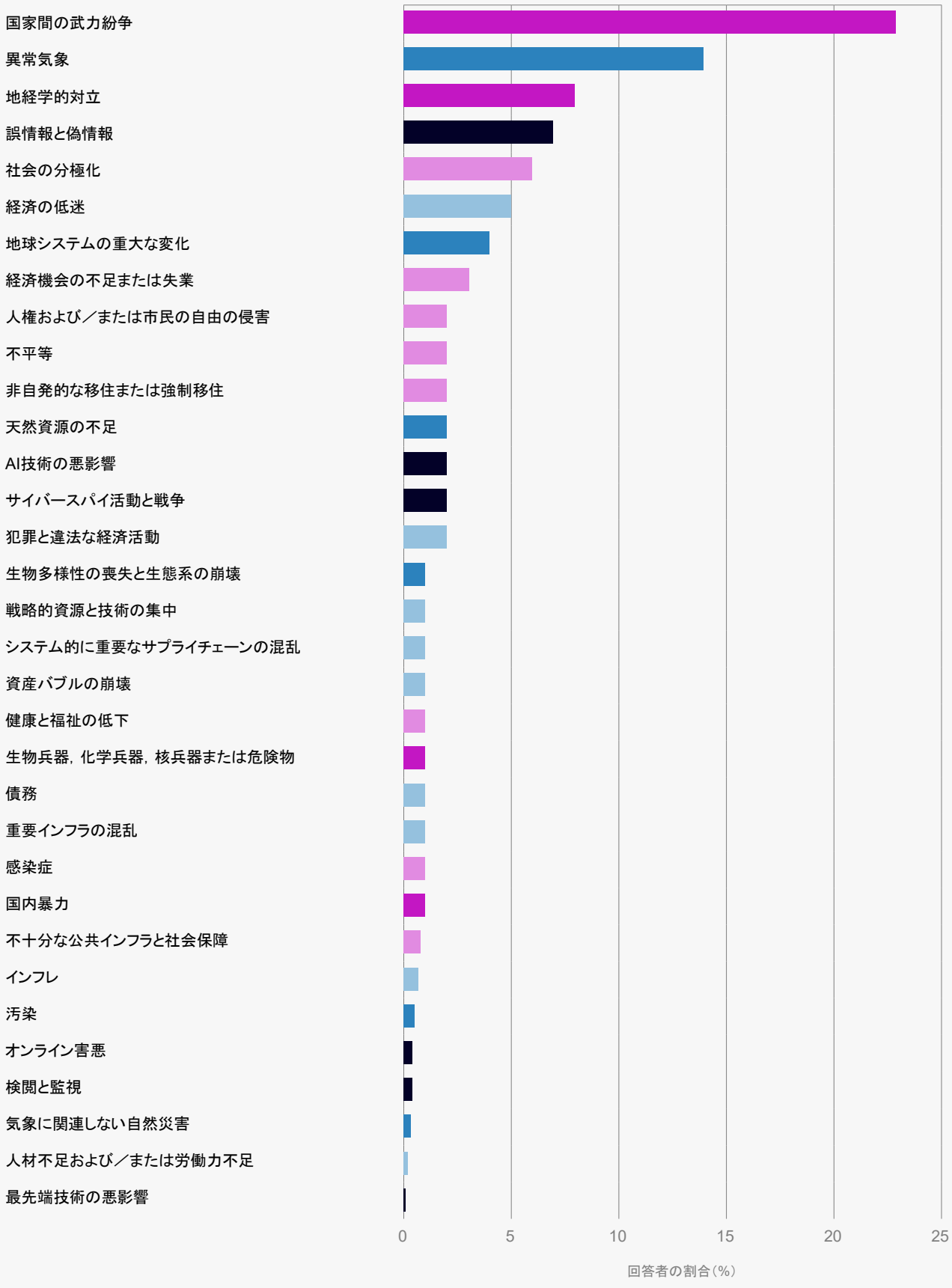
デジタルコンテンツの制作と配信の規模，スピード，そして高度化は，従来のコンテンツ検証手法を凌駕しています。コンテンツを検証し，知的財産を保護し，イノベーションを阻害することなく国民の信頼を維持するためには，新たなツールと戦略が必要です。

これらの課題とそれが社会に及ぼす影響により，誤情報と偽情報の問題は急速に公共政策のレベルへとエスカレートしました。世界中の政府は，規制手段，専門的規格，そして啓発キャンペーンを組み合わせて対応しています。

誤情報と偽情報は現在，世界で最も差し迫ったリスクの一つに挙げられています。世界経済フォーラムの「グローバルリスク報告書2025」によると，誤情報と偽情報は2年連続で世界最大のリスクとなっています。生成AIによって生成されるコンテンツの高度化が進むにつれ，真実と虚偽を見分けることがますます困難になっています。特に，合成メディアによって現実の経験と捏造された経験の境界が曖昧になっているため，その傾向は顕著です。



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



出典：世界経済フォーラム グローバルリスク報告書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

1.2　定義の重要性：誤情報，偽情報，悪意の情報

誤情報と偽情報はほぼ同義語となっていますが，特にその動機と適用方法においては，それぞれ明確に区別されています。議論の中で見落とされがちなのが悪意の情報です。フェイクニュースのコンテキストにおける悪意の情報は，虚偽を広めることを目的とした組織的なキャンペーンの一環として，偽情報と組み合わせて使用される場合，特に危険です。

- 誤情報とは，虚偽の情報を指しますが，害を及ぼす意図を持って作成または共有されたものではありません。[3]
- 偽情報とは，誤解を招いたり，害を及ぼしたり，操作したりするために意図的に作成および配布された虚偽のコンテンツです。
- 悪意の情報とは，事実に基づく情報をコンテキストから外し，害を及ぼす目的で利用することです。例えば，悪意を持って個人情報を公開すること（リベンジポルノや合意のない性的な画像など），あるいはコンテキスト上のメタデータを改ざんして誤解を招くことなどが挙げられます。

誤情報と偽情報の様々な種類をまとめた表は，附属書1に掲載されています。

> 虚偽または誤解を招くコンテンツの蔓延は，地政学的環境を様々な方法で複雑化させています。これは，外国勢力が有権者の意思に影響を与える主要な手段であり，紛争地域で何が起こっているかについて世界中の一般市民に疑念を抱かせ，あるいは他国の製品やサービスのイメージを傷つけるために利用される可能性があります。
>
> 世界経済フォーラム

プロパガンダ，詐欺，フェイクニュースといった手法は新しいものではありませんが，デジタル技術の発展により，よりアクセスしやすく，拡張性が高く，影響力の強いものとなっています。歴史的に戦争や政治の手段として利用されてきた偽情報は，今日では国家主体と非国家主体の双方によって利用され，難民，移民，社会的弱者といった脆弱な立場にある人々に壊滅的な影響を及ぼしています。

[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.

今日のハイパーコネクテッドなデジタル環境において，偽情報はまるで伝染病のように作用し，その急速な拡散は公共の言説を不安定にし，民主主義のレジリエンス（回復力）を蝕む恐れがあります。国内の主体であれ外国の主体であれ，虚偽の情報が組織的に展開されると，医療，気候変動政策，国家安全保障といった重要な分野における国民の信頼を損なう可能性があります。こうしたキャンペーンは，実証的な証拠に疑問を投げかけ，社会の分断を深め，複雑な地球規模の課題に対処するために必要な集団的合意の形成を困難にします。

1.3　どのような種類の加害者が問題を引き起こしているのか？

誤情報や偽情報を拡散する標的には，いくつかの種類があります。

個人：一般市民は，故意に（あるいは無意識に）有害なコンテンツを拡散させる可能性があります。意識的か否かに関わらず。ディープフェイクなどの技術により，説得力のある画像や音声を捏造することがかつてないほど容易になっています。

政治候補者および政治団体：候補者や政治団体は，虚偽の情報を用いて世論に影響を与え，分極化を深刻化させ，選挙の公正性を損なう可能性があります。

ソーシャルメディアプラットフォーム：これらのプラットフォームは，正確性よりもエンゲージメントを優先するため，虚偽の拡散を助長します。エコーチェンバー現象は既存の信念を強化し，影響を受けた人々に修正を届けにくくします。

モデル開発者および提供者：これらは，コンテンツの信憑性と帰属，誤情報の増幅，モデルの学習方法，使用するデータ，そして出力のモデレーション方法に関する透明性の欠如など，様々な課題を引き起こします。

従来のメディア：従来のメディアは，従来の安全対策を講じているにもかかわらず，特にデジタル時代においては，情報操作から逃れることはできません。ディープフェイクやモデレーションされていないユーザーコンテンツ（例：コメント欄）は，この問題をさらに複雑化させます。

国民国家／外国の主体：国民国家や外国の主体は，社会を不安定化させ，世論を操作するために，外国の悪意ある影響力による破壊活動の一環として，トロールファームやスポンサー付きのインフルエンサーといった組織的な偽情報戦略を用いる場合があります。

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

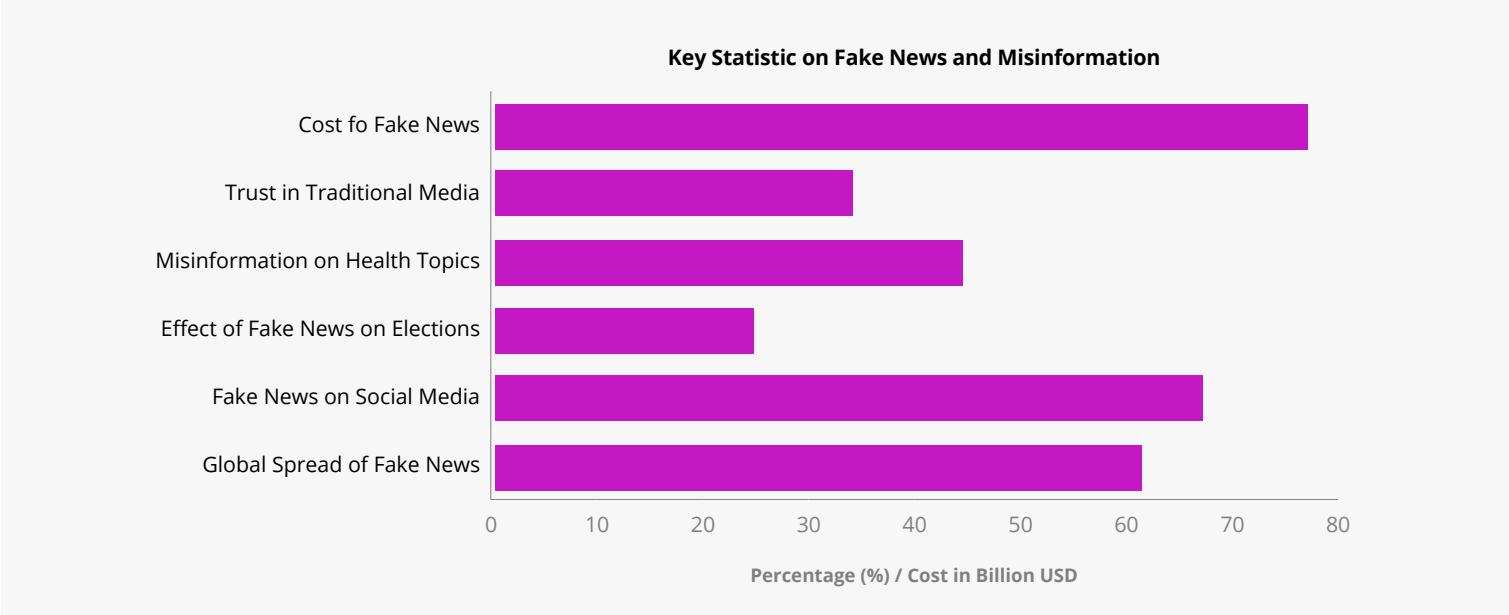


Figure 1: Key statistics on Fake News and Misinformation[4], Source SDLC Corp

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

これらの活動は，図1に示すように，多大な経済的および社会的コストを伴います。

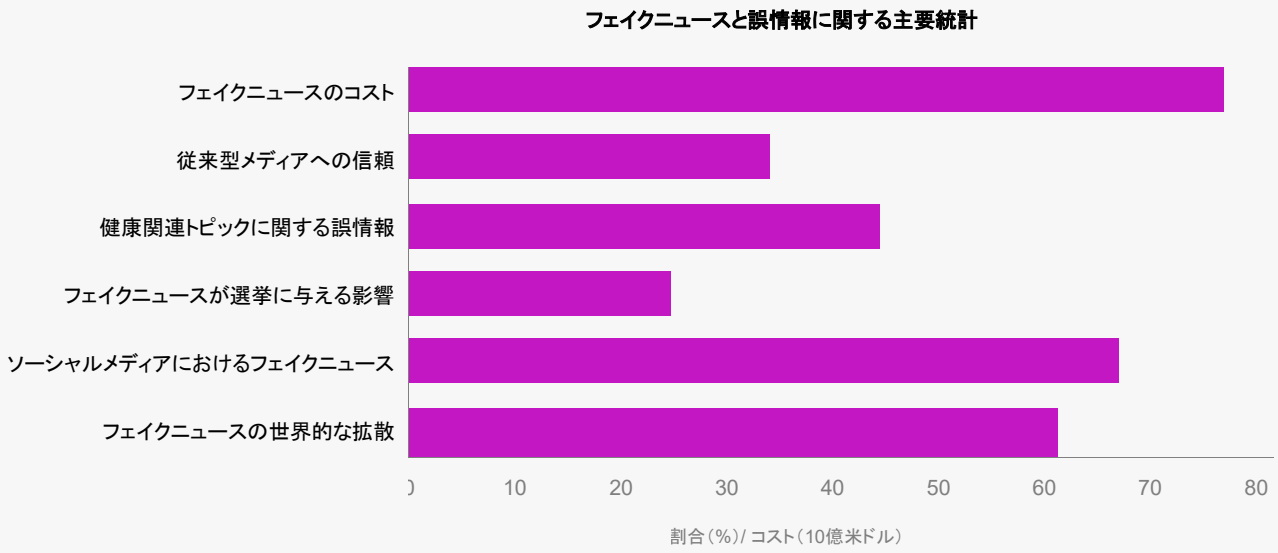


図1：フェイクニュースと誤情報に関する主要統計[4]，出典：SDLC Corp

1.4　ディープフェイクとサイバー攻撃

ディープフェイク[5]は，当初は娯楽や芸術目的で作成されましたが，現在では武器として利用されています。敵対者がリアルな画像，動画，音声録音を容易に偽造できること，そして合成コンテンツと非合成コンテンツの区別がますます困難になっていることは，サイバー犯罪者に高度な攻撃を仕掛ける絶好の機会を与えています。

「ディープフェイクと合成コンテンツの悪用は，国家安全保障，法執行，金融，そして社会のあらゆる分野において，国民にとって明白かつ現実的で，かつ進化し続ける脅威となっている。」[6]

米国国土安全保障省

超リアルな画像，動画，音声録音は，高度な詐欺，個人情報窃盗，ソーシャルエンジニアリング攻撃にますます利用されています。

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

[5] ディープフェイクのカテゴリーについては附属書3を参照

[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

金融セクターは特に脆弱です。最近のMediusの報告書によると，金融専門家の53%がディープフェイク詐欺の標的となり，そのうち43%が被害に遭っています。[7] 注目すべき事例として，ある金融従業員がディープフェイク動画を使って詐欺師に3,900万ドルを送金させられました。[8]

その影響は金銭的な損失にとどまりません。政治家，著名人，インフルエンサーなどの著名人は，深刻な風評被害に直面します。皮肉なことに，これらの人たちのキャリアを支えているメディア自体が，当人に不利なように操作される可能性があります。.

1.5 AIとマルチメディアの真正性

デジタルコンテンツ／マルチメディアへの信頼は，その完全性，出所，系譜，そして文脈が保持されているという信念の上に築かれます。これには，これらのメディアの制作者が，盗作，誤情報，偽情報を避けるための厳格な倫理規範に従っていることの確認も含まれます。コンテンツが同意なく改変された場合，特に法務，金融，ジャーナリズムの分野では，重大な影響が生じる可能性があります。

組織にとって，コンテンツライフサイクル（作成から管理，配信に至るまで）の整合性を確保することはますます困難になっています。誰がコンテンツを作成，変更，または消費したのか，そしてそれが依然として真実を反映しているのかどうかという疑問が生じます。品質とコンテンツガバナンスの基本基準を満たさないと，個人や組織は法的および規制（データ保護や知的財産権を含む）や評判リスクにさらされます。

贋作やメディア操作は，贋作絵画から写真の改ざんに至るまで，長らく存在してきました。今日の違いは，生成AIが現実を複製，偽造，または歪曲できる規模と速度にあります。例えば，1800年代の心霊写真やエイブラハム・リンカーンの加工された肖像画は，恐ろしいほどの精度で声や身元を偽装できる今日のディープフェイクとは比べものになりません。

これは被害者を危険にさらすだけでなく，あらゆる形態のメディアに対する国民の信頼を損ない，信頼できるメディアを完全に排除することにつながります。これは，特に正当な証拠が捏造されたと誤って認識された場合，ジャーナリズム，ガバナンス，司法，そして社会の結束に深刻な影響を及ぼします。リスクはもはや騙されることだけではありません。真実を含め，あらゆるものに対して懐疑的になることです。

生成AIによって合成情報と非合成情報の境界が曖昧になるにつれ，個人が見たり，聞いたり，読んだりするものを信頼することが難しくなります。この信頼性の危機は，政府から企業，ジャーナリスト，教育者，そして一般市民に至るまで，あらゆる人々に影響を与えます。それぞれのグループは，露出度と脆弱性に応じて，さまざまなレベルの影響を経験するでしょう（ステークホルダーへの影響については附属書2を参照）。コンテンツへの信頼が低下するにつれて，リスクは法的，社会的，倫理的，そして専門的な領域にまで及びます。誤情報の拡散が拡大し，情報源の信頼性が低下するにつれて，私たちは専門的，倫理的，そして社会的な懸念にまたがる複雑な課題に直面しています。必要なのは，コンテンツ検証における緊急のイノベーションと，健全な法的・規制的フレームワークと国際規格によって支えられるデジタルリテラシーの向上です。

[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.

第2章

マルチメディアの真正性に対する信頼を構築するために，規制環境と市場ニーズのバランスを取る複雑さ

2013年，世界経済フォーラムは「オンライン上の誤情報の急速な拡散」を10大グローバルリスクの一つに挙げました。10年以上経った今でも，この懸念は依然として最前線にあります。同フォーラムは「グローバルリスク報告書2025」において，誤情報と偽情報が世界で最も差し迫った課題の一つであることを改めて強調しました。

> 誤情報と偽情報は2年連続で短期リスクのトップに挙げられ，信頼を損ない，国家間および国家間の分断を悪化させることで，社会の結束とガバナンスに対する根強い脅威となっていることを浮き彫りにしています。
>
> 世界経済フォーラム

度重なる警告と，経済的，社会的，そして風評被害の拡大にもかかわらず，依然として疑問が残ります。なぜこの問題は依然として存在するのでしょうか？

2.1　マルチメディアの真正性に対する信頼構築はなぜ複雑なのでしょうか？

マルチメディアの真正性に対する信頼構築は，そのエコシステムの相互依存的な構成要素と，幅広いステークホルダーが関与していることから，本質的に困難です。この問題をさらに複雑にしているのは，世界的に認められたデジタルIDフレームワークが存在しないことです。そのため，特に国境を越えた場合には，個人または組織のアイデンティティを確実に検証することが困難になっています。その結果，ID窃盗，なりすまし，合成IDに対する脆弱性が高まっています。

信頼を獲得するには，以下が必要です。

- 包括的かつ一貫性のあるフレームワークを確立する明確な国際および国内政策と規制
- これらのフレームワークを一貫して適用するための，セクター全体にわたる組織的なコンプライアンスとサポート
- 規制要件に沿って設計・導入された技術的ソリューション
- 一貫性と意義のある実施を確保するための，強制力と自主性の両方を備えた強固な執行メカニズム

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.

2.2　困難をもたらす要因とは？

これを達成することの困難さには，いくつかの要因があります。

- これは世界的な問題ですが，実施と執行は国レベル（場合によっては地方レベル）で行われ，多くの場合，様々な政治理念や管轄上の制約，そして市場の要件によって影響を受け，形作られます。例えば，違反に対する罰則と国境を越えた執行措置について合意を形成することが不可欠です。
- この問題は，消費者保護，知的財産，国家安全保障など，複数の分野にまたがるため，単一の規制でマルチメディアの真正性の全容に対応することはできません。EUの一般データ保護規則（GDPR）は，データプライバシーに焦点を当てていますが，他の分野にも参考にできる優れた基盤を提供しています。GDPRはEUとイギリスを対象としているにもかかわらず，域外適用が可能であり，その結果，同様の問題に直面している他の国々で同様の法律の審議が始まりました。
- 政策立案者は，オンライン上の危害を防ぎながら，表現の自由を保護し，イノベーションを促進し，投資を誘致するなど，相反する優先事項のバランスを取る必要があります。
- 規制の効果的な実施は，規制に準拠した技術ソリューションの開発を含む，業界からの強力な支援と協力にかかっています。
- 規制の能力と成熟度は国によって異なります。各国の規制の策定，実施，執行能力には大きな差があり，世界的な整合性と調整が大きな課題となっています。
- 技術の急速な進歩と規制の遅れの間には緊張関係があります。生成AI，クラウドコンピューティング，国境を越えたデータフローの急速な進化は，規制システムのペースを上回っています。データの所在に関する管轄権の曖昧さと，国際的なインターネット機関の不在は，分断をさらに悪化させています。

2.3　規制環境の概要

オンラインの安全性，誤情報，偽情報の規制に向けた取り組みは，国際原則，国際ガイドライン，そして各国の規制フレームワークの組み合わせによってますます強化されています。これらには，政府，国際機関，市民社会，テクノロジー企業からの貢献が含まれており，多くの場合，安全性の確保，透明性，説明責任といった概念が重視されています。

誤情報は様々な分野やプラットフォームにまたがっているため，世界中で多様な法的・政策的メカニズムが生まれています。以下は，主要な国際的な取り組みの概要と，それに続く地域および国の規制フレームワークです。

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

国際的な取り組み

- OECDデジタルサービスプロバイダーガイドライン：特に子供や脆弱なユーザーを保護するために，リスクに基づくアプローチを推進しています。
- テロ対策のための世界インターネットフォーラム（GIFCT）：有害コンテンツの検出と削除を目的とした官民パートナーシップです。
- Tクライストチャーチ・コール：ニュージーランドとフランスが主導するこのイニシアチブは，政府とテクノロジー企業が協力して，テロリストや暴力的過激主義のコンテンツを排除することを目指しています。
- GPAIとOECDのイニシアチブ：コンテンツモデレーションと情報の完全性におけるAIの責任ある活用を促進します。
- ユネスコデジタルプラットフォーム規制ガイドライン（2023年）。国連ビジネスと人権に関する指導原則を補完し，誤情報と偽情報に対処するための人権に基づく原則を概説しています。
- ユネスコ人工知能倫理勧告：2021年に採択され，ユネスコ加盟国194か国すべてに適用され，政策行動分野に関する勧告を行っています。
- ユネスコメディア情報リテラシー（MIL）フレームワーク：ユーザーが情報の信頼性を批判的に評価できるようにし，世界的なデジタルリテラシーの向上を目指しています。

2022年に設立されたグローバルオンラインセーフティ規制当局ネットワークは，国際的なオンラインセーフティ規制当局（オーストラリアのeSafety Commissioner，イギリスのOfcom，アイルランドのCoimisiún na Meán，フィジーのオンラインセーフティ委員会，韓国の放送通信委員会（BCC），その他ヨーロッパ，北アメリカ，アジア太平洋地域の規制当局を含む）の連合体です。ネットワークの目的は，以下のとおりです。

- 協力を通じて安全なオンライン環境を促進する
- エビデンスに基づく政策策定を支援する
- 画一的な解決策を強制することなく，規制アプローチの整合性を促進する

オンラインセーフティ規制指標[9]は，さまざまな法域におけるオンラインセーフティ規制への取り組み方を比較分析し，以下の情報を提供しています。

- 各国の立法モデル
- 執行の成熟度
- 共通原則（例：児童保護，システミックリスク）
- 実践における世界的な動向と収束／発散

これは，政策立案者が世界的な動向を把握し，プラットフォームが管轄区域をまたいでコンプライアンスを遵守できるよう支援し，規制間の相互運用性を促進するのに役立ちます。

[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.

地域および国レベルのアプローチとフレームワーク

欧州連合：

- 一般データ保護規則（GDPR）2018年
  - プライバシーに焦点を当て，データの不正使用や，誤情報（マイクロターゲティングなど）を助長する可能性のあるアルゴリズムによるプロファイリングも制限
- デジタルサービス法（DSA）2022年
  - オンラインプラットフォームに対する包括的な規制
  - コンテンツモデレーションの透明性，アルゴリズムの説明責任，偽情報などのシステムリスクの軽減を義務付け
  - 大規模オンラインプラットフォーム（VLOP）に厳格な規則を適用
  - 偽情報およびヘイトスピーチへの迅速な対応を義務付け
- 偽情報に関するEU行動規範（2022年改訂）
  - Meta，Googleなどの主要プラットフォームが署名した，自主的ながらも制度化が進む規範
  - 政治広告の透明性，虚偽コンテンツの収益化停止，ファクトチェッカーへの支援を義務付け
- 「人工知能と人権，民主主義及び法の支配に関する欧州評議会枠組条約」は，以下の点について勧告を行っています。
  - ファクトチェック
  - プラットフォーム設計ソリューション
  - ユーザーのエンパワーメント
- EU AI法
  ディープフェイクに対処するため，EU AI法は第50条（2）に基づき透明性を促進している。汎用AIツールの提供者に対し，AI生成コンテンツにタグを付け，改ざんを特定することを義務付け，ユーザーが情報をより深く理解できるようにします。ただし，これは軽微な修正などの標準的な編集作業，または犯罪捜査や訴追などの法執行活動のために許可されている場合には適用されません。

  EU AI法，特に前文133は，多様なコンテンツ形式，検出方法，AI機能に対応するための柔軟性の必要性を認めています。これにより，特に多様なコンテンツや進化する技術を扱う事業者にとって，効率的なコンプライアンスが確保されます。前文133はさらに，コンテンツの出所を追跡し，真正性を証明するための透かし，メタデータタグ，フィンガープリント，セキュリティ機能といった技術を含む，正確で互換性があり効果的なタグ付けおよび識別ツールの重要性を強調しています。重要な懸念事項として，ディープフェイクの分類に関する曖昧さが挙げられます。EU AI法はAI生成コンテンツの開示を義務付けている一方で，欺瞞的なディープフェイクを高リスクと明示的には指定していません。

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.

アフリカ:

- サイバーセキュリティ及び個人データ保護に関するアフリカ連合条約（マラボ条約），2014年
  - アフリカ連合加盟国に対し，データプライバシーと表現の自由を保護するサイバー犯罪対策法の制定を奨励しています。
- 各国の例（ナイジェリア，ケニア，南アフリカなど）
- 言論の自由に対する懸念は依然として残るものの，サイバー犯罪やヘイトスピーチに関する法律を用いて偽情報への対処を図ることが多いです。

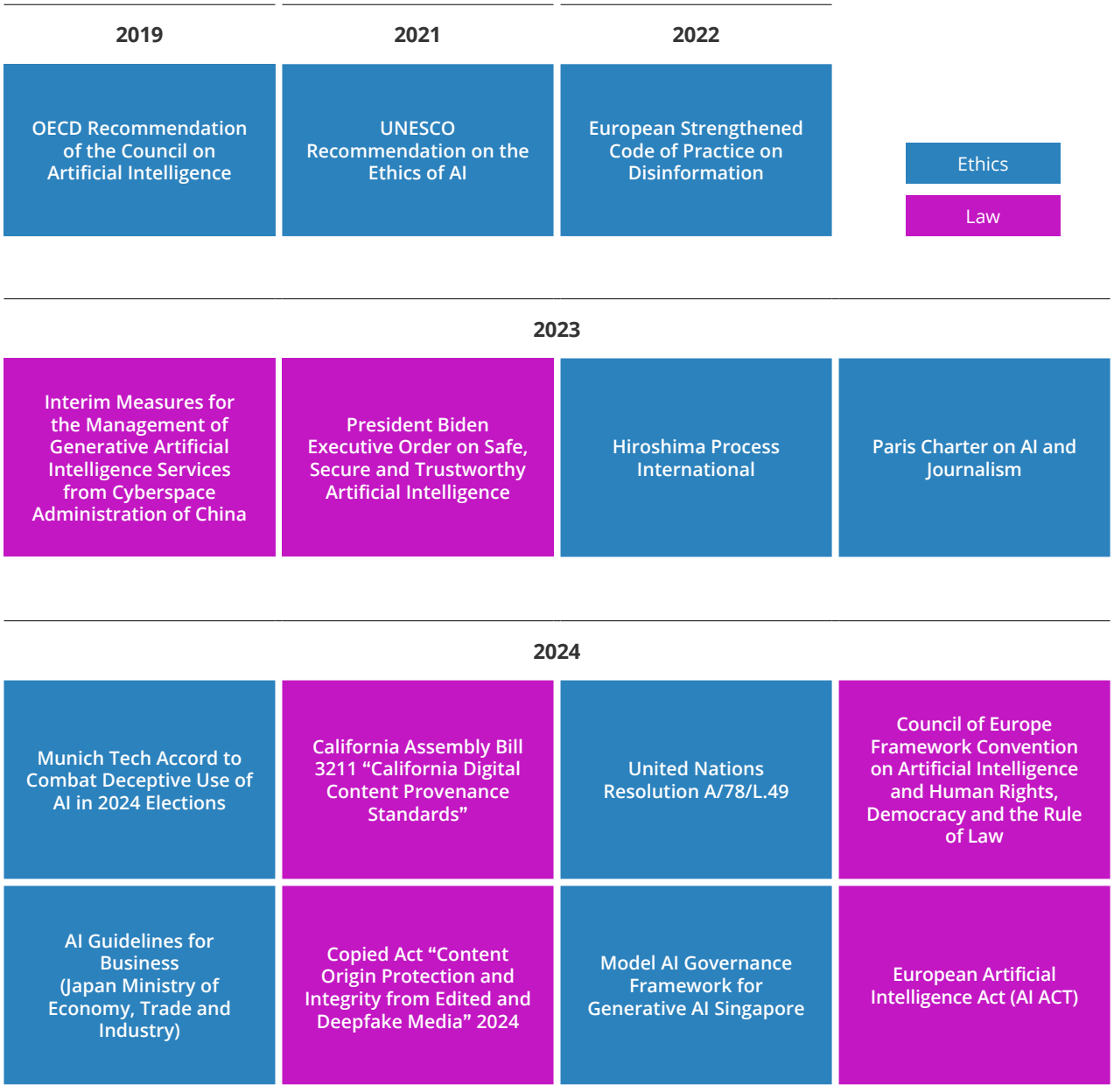
アジア太平洋地域:

- ASEANデジタルマスタープラン2025は，東南アジア全域におけるデジタルセーフティ協力とメディアリテラシーの向上を推進しています。
- オーストラリア：オンラインセーフティ法（2021年）は，eセーフティ・コミッショナーに有害コンテンツを削除する権限を与えています。設計段階からの安全性を重視し，ネットいじめや偽情報から保護します。
- インド：IT規則（2021年）は，コンテンツの迅速な削除，発信者の追跡を義務付け，「重要なプラットフォーム」に対してより厳格な規則を課しています。
- シンガポール：オンライン虚偽情報および情報操作防止法（POFMA）（2019年）は，政府による訂正命令の発令や虚偽コンテンツへのアクセスブロックを認めています。これは，言論の自由への影響の可能性について批判を受けています。
- 中国：深層合成インターネット情報サービス管理に関する規定（2023年）は，生成AIおよびディープフェイク技術の使用を規制しています。この規定は，プラットフォームに対し，AI生成コンテンツにラベルを付け，悪用を防止し，合成メディアが虚偽または有害な情報を拡散しないようにすることを義務付けています。

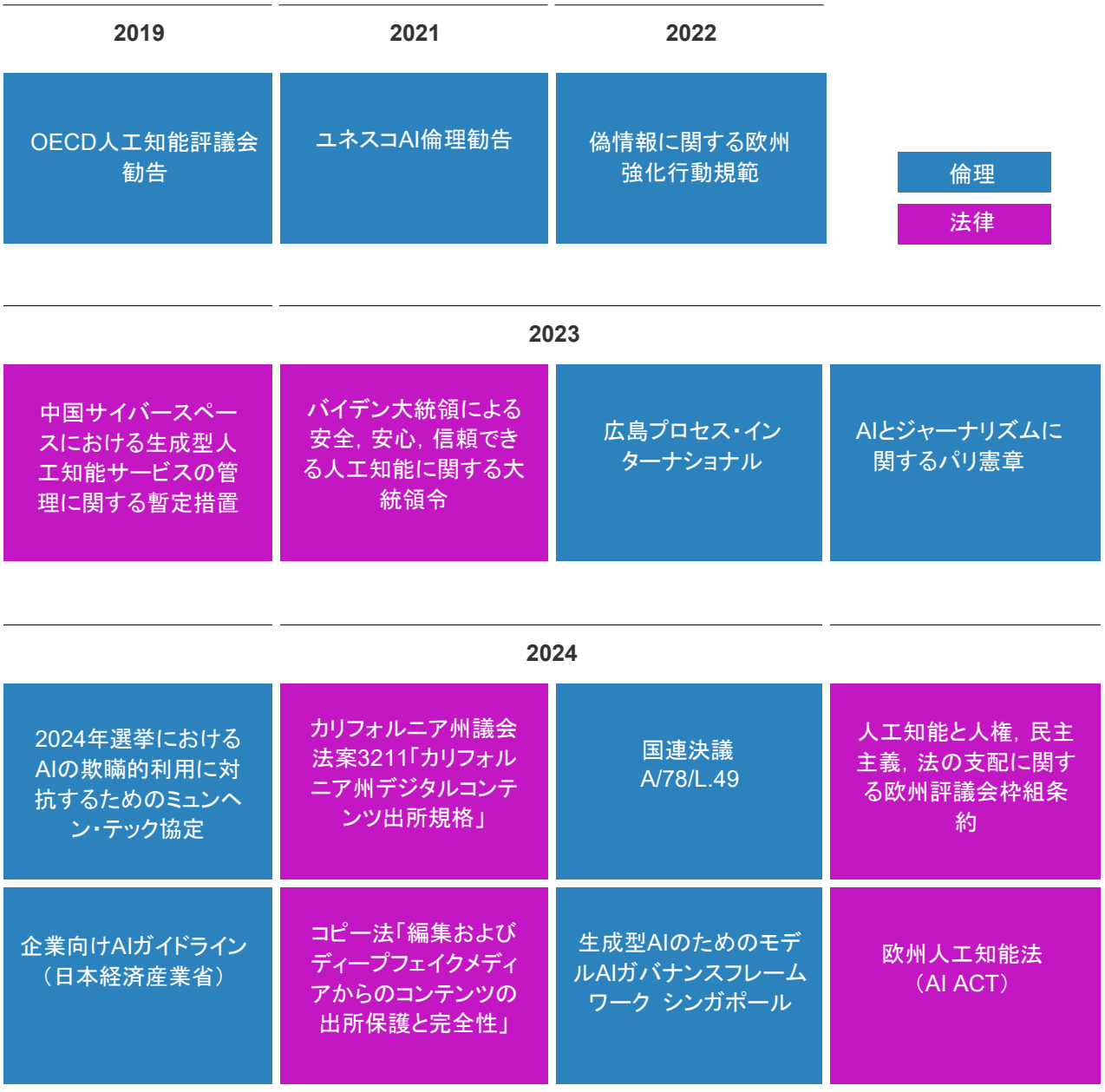
南北アメリカ:

- アメリカ合衆国：憲法修正第一条の保護により，誤情報に関する包括的な連邦法は存在しません。主な要素は以下のとおりです。
  - 通信品位法第230条：プラットフォームの免責を規定するとともに，モデレーションを可能にします。
  - FTCによる執行：偽情報に関連する欺瞞的な商慣行を標的としています。
  - 州レベルの取り組み：例：カリフォルニア州年齢相応設計法は，子供の安全対策に取り組んでいます。
- カナダ：
  - オンライン危害法（法案C-63，2024年，提案）：ヘイトスピーチや誤情報を含む有害なオンラインコンテンツを規制することを目指しています。
  - デジタル市民イニシアチブ：偽情報対策のための教育と研究に資金を提供します。
- ブラジル：フェイクニュース法案（PL 2630，提案）：この法案は，ユーザーIDの確認，バイラルメッセージの追跡，スポンサーコンテンツの開示を義務付け，特に選挙や健康に関する誤情報に対抗することを目的としています。
- イギリスオンライン安全法（2023年）：この法案は，プラットフォームに対し，違法で有害なコンテンツ，特に子供に影響を与えるコンテンツに対処するための注意義務を課します。Ofcom（イギリス情報通信庁）の規制を受け，広範な社会的影響を及ぼす誤情報に関する規定が含まれています。

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:



国際規格は，イニシアチブや連携と相まって，規制当局間の連携を実現するための強力なメカニズムを形成し，ユーザーの信頼を構築し，AIを活用したマルチメディア技術の安全な導入を可能にするために不可欠です。以下の図は，コンテンツラベリングに関する倫理的および法的フレームワークの進展を示しています。



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

2.4　規制と信頼のギャップの橋渡し

政策立案者や規制当局が直面する大きな課題の一つは，生成AIのようなマルチメディアの真正性は，特にオンライン安全規制の観点において，根本的に「ブラックボックス」であるということです。これらのモデルの開発とトレーニング方法に関する透明性は限られています。技術は大きな可能性を秘めていますが，根底にある運用が概ね不透明な状況において，効果的なガバナンスをどのように実現するかという問題が浮上しています。イノベーションを阻害することなく，マルチメディアコンテンツの信頼性と解釈可能性を確保するという主要な課題は，より広範な懸念事項と重なり合っています。これには，誤情報対策といった新たな世界的な優先事項との整合性，そしてオンライン安全規制によってそれらの優先事項がどのように形成され，影響を受けるかといった問題が含まれます。

グローバルオンラインセーフティ規制当局ネットワークは，最初の年次報告書[10]および2025～2027年戦略計画[11]において，以下のテーマを重点課題として挙げています。

- 法域間の規制の一貫性の構築
- オンラインセーフティに関するエビデンスベースの構築とベストプラクティスの共有
- コンプライアンス促進のための情報共有と連携の促進

現在，主要なオンラインセーフティ対策の状況と適用，そしてそれらが対処するオンライン被害の種類について，混乱が生じており，明確性に欠けています。これは，誤情報や偽情報に関するリスク軽減の実現に大きく影響します。国境の不在を悪用できてしまう技術の性質上，各地域の取り組みが可視化されなければ，公平で認識可能なガバナンス体制の確立は困難です。規制の地域的適用範囲，各法域がどのように規格や法律を運用しているか，そして上記のようにそれぞれの状況を明確に把握することは，現在もなお課題であり，今後も続くでしょう。

適用可能かつ適切な国際規格に基づき，将来を見据えた，技術の発展に合わせて進化できる合意された政策と規制のフレームワークを構築することは，多様な関係者の協力を必要とする広大な課題です。オンライン安全対策は，あらゆるマルチメディア利用分野と，それによってもたらされる危害に対する包括的な対策において不可欠な要素です。連携がなければ，遡及的に埋めるのが困難なアプローチのギャップが生じるリスクがあります。

しかし，誤情報，偽情報，ディープフェイク，そしてマルチメディアの真正性に対する各国のアプローチの差を埋め，連携を図ることは可能です。この課題の規模を過小評価する人はいません。解決策を見出すという共通の使命を持つ多くの共同プロジェクトが進行中であり，その努力の規模の大きさを物語っています。

[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.

これらの共同プロジェクトは，規制当局や執行機関だけではマルチメディアへの信頼を築くことはできないという，複数のステークホルダーの認識を示しています。すべての関係者が協力し，新たな形の国際協力と規制，さらには自主規制を模索する必要があります。これには，信頼を育み，人権，メディアリテラシー，そして個々の利用者の倫理観を含む企業責任も組み込む必要があります。

しかしながら，実行可能で適用可能な解決策を開発することなく，常に問題について議論を続ける限り，関係者間の協力や取り組みの推進を求めることは，主に哲学的な傾向にとどまってしまうでしょう。

次の章で議論するように，これらの取り組みを展開する一つの方法は，既存のフレームワークや規格を基盤とし，政府や産業界が採用できる実用的なソリューションを提案することです。

2.5　政府と産業界のための実用的なソリューションの創出

上記で強調した課題の多くは，マルチメディアの真正性に対する信頼を構築するために，各国政府が予防-検知-対応（PDR）フレームワークをどのように導入しているかを検証することで，より深く理解し，対処することができます。この3本柱のアプローチは，規制の意図と技術的な実現可能性のバランスを取りながら，拡張性と柔軟性に優れた構造を提供します。

表1. MMAへのPDRフレームワークの適用

アプローチ	政策要件	方法	利点および/または成果
防止	透明性	ラベリング	コンテンツのさまざまな側面についてユーザーに通知する。コンテンツがAIによって生成されたかどうかを明確に識別します。
		透かし	コンテンツに適用される，人間には認識できないマーキングで，コンテンツに関する情報を提供します。
	トレーサビリティ	コンテンツ出所ツール	コンテンツの出所と変更に関する情報を提供することで，説明責任と帰属を明確にします。
	説明責任	リスク評価の実施	高リスク地域が特定された場合，執行をより効率的に行うことができる。蔓延している不正行為や行動パターンを特定し，優先的に対処します。この積極的なアプローチは，操作されたコンテンツに関連するリスクを軽減し，ユーザーを誤情報や詐欺行為から保護するのに役立ちます。
	ユーザー教育	啓発活動	著作権法と侵害の結果に関する教育を通じて，偶発的な誤用を削減します。

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

アプローチ	政策要件	方法	利点および/または成果
検出	操作されたコンテンツとディープフェイクの検出	技術的ソリューション	これらのソリューションは，知的財産の保護，画像，音声，テキスト，動画の真正性の検証，オンラインの安全性とセキュリティの向上など，多くの利点をもたらします。しかし，これらの検出機能を回避しようとする悪意のある行為者との「攻防戦」を引き起こします。 例: https://arxiv.org/abs/2504.2148
	データプライバシー	データの取り扱いとデータ保護法の遵守	処理されるすべてのデータは，ランダム化された手動レビューの対象となり，正確性とデータ保護法の遵守を確保します。
対応	執行	規制介入	政府が，使用すべき技術とアプローチを具体的に規定する法律や規制を制定することにより，罰則を適用し，規則を施行することができます。また，これらの技術が侵害された場合の対応についても規定しています。
	説明可能性	説明型アルゴリズム，AIモデル検証手法，および使用されるトレーニングデータセットに関する情報の使用	AIシステムによる決定は，高い信頼性を維持するためにチェックすることができます。これは，知的財産権侵害のリスクを軽減するのに役立ちます。
	紛争解決メカニズム	コンテンツの争訟性	明確で適切に伝達されたメカニズムは，個人にとって有益であり，主張への異議申し立てに役立ちます。
		プラットフォームの禁止	問題のある地域の取り締まりは，侵害コンテンツを頻繁にホストするプラットフォームやウェブサイトへのアクセスを定期的に削除することで，より効果的かつ有益になります。

このフレームワークは，プライバシー（GDPR，CCPAなど）やサイバーセキュリティ（NISTサイバーセキュリティフレームワーク[12]，PCI-DSSなど）における成功したアプローチを反映しています。PDRの強みは，そのシンプルさと汎用性にあります。PDRは広く理解されており，あらゆるセクターに適応可能で，規制の整合性にも役立ちます。プライバシーの場合，成功したアプローチは，予防（プライバシー・バイ・デザイン），検知（侵害の通知と監視），対応（執行措置とユーザー救済メカニズム）を重視しています。これらの規制は主にプライバシーに焦点を当てているように見えますが，データ（特に部分的または完全に合成されたコンテンツ）の使用方法と伝達方法に関する明確で透明性が高く正確な情報の重要性を強調することで，マルチメディアの真正性に対処するための貴重なモデルを提供しています。

[12] NISTのサイバーセキュリティフレームワークは，予防，検知，対応に加えて，識別と回復の機能も備えています。
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.

しかし，PDRフレームワークを適用するには，専門的な視点だけでなく，社会技術的な視点も必要です。これには，予防，検知，対応の各段階において，人間の行動と倫理的利用，ビジネスプロセスと市場インセンティブ，そして技術の設計と導入の間の複雑な相互作用を認識することが含まれます。

PDRベースのフレームワークを組織レベルで導入することで，規制の整合性，一貫性，そして公平なコンプライアンスの達成可能性が高まります。この共通構造は，規制の導入を促進し，説明責任を促し，政府と市場関係者間のコミュニケーションを円滑化するのに役立ちます。

さらに，PDRは規制の執行可能性を高めます。公共部門と民間部門の両方が共通の構造で運営されることで，規制目標はより実行可能になります。まさにこの点において，国際規格と適合性評価はPDRの導入において重要な役割を果たします。

次の章では，政策と技術のギャップを埋める上での国際規格の役割を考察し，マルチメディアの真正性に対する信頼を支える具体的な規格を概説します。結局のところ，あらゆるPDRフレームワークの有効性，特にマルチメディアの真正性のような複雑な分野においては，関連する国際規格との整合性が重要です。これらの規格は，PDRの各柱を支えるために必要な専門的および手続き的な基盤を提供します。

PDRのような実績のあるモデルを将来の規制の基盤とし，あらゆるレベルで規格を組み込むことで，関係者はより信頼性が高く，回復力のあるデジタル情報エコシステムを共同で構築することができます。

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.

第3章

マルチメディアの真正性への対応における国際規格と適合性評価の役割

生成AIの急速な進化，マルチメディアの作成，編集，および／または操作への影響の増大，そして誤情報や偽情報の拡散の増加は，世界中の政府や規制当局にとってますます複雑な課題を突きつけています。これらのリスクに効果的に対処するには，協調的かつ調和のとれた行動が不可欠であり，特にマルチメディアの真正性を検証するためのメカニズムの相互承認を可能にする規格の開発が不可欠です。これがなければ，国境を越えた規制のギャップが解消されず，分断，非効率性，脆弱性につながるでしょう。

国際協力は効果的な対応の礎です。国際規格，適合性評価手順，そしてより広範な品質インフラ（QI）システムが，この連携を支えるべきです。[13] これらのツールは，今日の課題に対処するために必要な専門的およびガバナンスのフレームワークを提供するだけでなく，急速な技術発展に合わせて規制が進化することを確実にします。

コンテンツの真正性の相互承認を促進し，国境を越えた抜け穴を塞ぐために，各国政府は国際的に認められたコンセンサスに基づく規格を採用し，参照すべきです。国際規格は，政策立案者に次のような利点をもたらします。

- 法域間の相互運用性を支える共通の語彙とベンチマーク
- 安全性，セキュリティ，ガバナンス，アカウンタビリティに関する評価方法とベストプラクティスのフレームワーク
- オープンで柔軟性と適応性に優れたソリューションを推進することで，技術的な「ロックイン」または「ロックアウト」を回避するメカニズム

国際規格との整合性と，適合性評価の活用方法に関する明確な合意がなければ，規制の分断，作業の重複，そして官民の資源の非効率的な配分のリスクは増大する一方です。すべての規格が同等に作られているわけではないことに留意する必要があります。透明性が高く，多様なステークホルダーによるプロセスを通じて策定された，国際的に合意された規格を優先すべきです。

[13] INetQIによる定義：製品，サービス，およびプロセスの品質，安全性，環境健全性を支えるために必要な組織（官民），政策，法的フレームワーク，およびプラクティスを含むシステム。市場の機能を支え，外国市場へのアクセスを促進する包括的なフレームワークです。

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

3.1 国際規格の価値

ISO，IEC，ITUによって開発され，世界規格協力（WSC）として知られる国際規格は，市場ニーズに応え，多様な専門家の合意を反映するグローバルなツールです。包括的なマルチステークホルダー・プロセスを通じて開発されたこれらの規格は，社会，環境，技術，経済の側面に対応しています。

OECDの適正規制慣行（Good Regulatory Practices）と世界貿易機関（WTO）の貿易の技術的障害（TBT）協定は，どちらも規制における国際規格の活用を推奨しています。これらの規格は，WTO TBTの国際文書策定のための6原則に準拠しており，貿易に不必要な障害を生じさせず，規制協力を可能にするものとされています。規制，政策，または適合性評価制度において参照される国際規格は，次のような効果をもたらします。

- 既成のベストプラクティスを提供することで規制負担を軽減します
- 技術的ルールを政治サイクルから切り離すことで政策実施を迅速化します
- 国際協力を促進し，世界の貿易フローを円滑にします

公共政策，特に国連の持続可能な開発目標（SDGs）の推進において，国際規格は透明性，予測可能性，そして説明責任を強化します。国際規格は，持続可能な経済成長を促進しながら，費用対効果が高く効率的な政策実施手段を提供します。

規格がイノベーションを阻害するという一般的な誤解とは裏腹に，多くの研究が，十分に整備された国際規格がイノベーションを支え，推進することを示しています。国際規格は，研究開発のための安定した基盤を提供し，相互運用性を促進し，重複作業を削減することで，イノベーターが差別化された付加価値の高いソリューションの提供に集中することを可能にします。これは，透明性と互換性が市場投入までの時間を短縮する，ペースの速い競争の激しい環境において特に重要です。

これらの利点は，既存および新興の国際規格をPDRの取り組み全体を通して活用すべき多くの理由の一つです。本政策文書の後半では，関連標準規格とPDRのマッピングを示し，標準規格を実際にどのように適用できるかを示します。堅牢でコンセンサスに基づく標準規格に基づいた技術を開発・推進することで，政府と産業界は，信頼性，拡張性，イノベーションが相反するのではなく，互いに補完し合う関係を築くことができます。このような理由から，標準の必要性と適用可能性は，イノベーションへのプラスとマイナスの影響に関する研究を絶えず促しています。[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.

例：医療における国境を越えた導入

医療のグローバルな性質は，規格の有用性の好例です。例えば，ISO 14971:2019 医療機器 - リスクマネジメントの医療機器への適用 は，以下のように採用されています。

- アメリカ合衆国：ANSI/AAMI/ISO 14971
- 欧州：EN ISO 14971
- 日本：JIS T 14971

こうした協調的な採用は，世界的な規制の整合性を支援し，患者の安全を確保しながら貿易を促進します。

同様に，マルチメディアにおける国際規格は，次のようなことを可能にします。

- 倫理的なAI導入を導きます
- 出所および真正性プロトコルを定義します
- 検証済みのデジタルコンテンツを通じて公共の信頼を守ります

3.2 AIとマルチメディアの真正性：標準化の実践

マルチメディアの真正性に関する5つの主要分野において，国際規格は特に重要です。

1. コンテンツの出所
2. 信頼性と真正性
3. 透かし
4. 資産識別子
5. 権利宣言

「AIとマルチメディアの真正性規格に関する技術報告書：標準化の現状展望」は，デジタルメディアの真正性と人工知能に関する規格および仕様の現状を5つのクラスターに分けて包括的に概観しています。本政策文書では，前述の報告書で取り上げられた5つのクラスターのうち，コンテンツの出所，信頼性と真正性，透かしの3つに焦点を当てています。これは，これらのクラスターが誤情報と偽情報によって生じる問題に最も関連しているためです。

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

特筆すべき点は，2つの独立したマッピング作業（1つは社会技術／政策，もう1つは専門的）から得られた結果が重複しており，分野横断的な強いコンセンサスが示されたことです。

コンテンツの出所

規格番号	担当グループ	名称
ISO 22144	ISO TC 171/SC2	コンテンツ認証情報
ISO 21617-1:2025	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust 第1部
	オリジネータープロファイル	Originator Profile
	Open Provenance	PROV
	C2PA	コンテンツ認証情報
	DIFの一部である作成アサーション作業グループ	CAWGメタデータ

情報の信頼性と真正性

規格番号	担当グループ	名称
未定	ITU-TSG13 – ISO/IEC JTC 1/SG 29	H.MMAUTH: マルチメディアコンテンツの認証のためのフレームワーク
ISO/IEC TR 24028:2020	ISO/IEC JTC 1/SC 42	情報技術 – 人工知能 – 人工知能における信頼性の概要
ITU-T Y.3054	ITU-T	信頼に基づくメディアサービスのためのフレームワーク
JPEG Trust 第2部	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust 第2部
ISO/CD 22144	ISO	情報の真正性 – コンテンツ認証情報s
	DIF の一部である作成アサーション作業グループ	CAWGメタデータ
	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)

透かし

規格番号	担当グループ	名称
ISO/IEC 23078-1:2024	ISO/IEC JTC 1/SC 34	情報技術－デジタル出版物のためのDRM技術の仕様 第1部：出版業界で使用されている著作権保護技術の概要
SMPTE ST 2112-10:2020	SMPTE	コンテンツ識別子のオープンバインディング（OBID）
JPEG Trust 第3部	ISO/IEC JTC 1/SC 29/WG 1	JPEG Trust 第3部
2413-PLEN	ITU-T SG17	X.ig-dw：電子透かしの実装ガイドライン
ISO/IEC TR 21000-11:2004	ISO/IEC JTC 1/SC 29/WG 11	情報技術－マルチメディアフレームワーク（MPEG-21）－第11部：継続性連合技術の評価ツール
IEEE P3361	IEEE	デジタルコンテンツにおける電子透かし実装の堅牢性評価方法に関するIEEE規格原案
	NIH	遠隔放射線診断における医用画像透かし要件のレビュー
TR 104 032	ETSI	人工知能のセキュリティ確保（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.

その他の関連規格

AI生成マルチメディアへの信頼を築くには，AIバイアスが回避され，リスクが十分に考慮され，マネジメントシステムが要件を満たしていることを表明する必要もあります。国際的に認められた規格もここで重要な役割を果たします。

規格番号	担当グループ	名称
ISO 24027:2021	ISO	情報技術－人工知能（AI）－AIシステムとAI支援の意思決定におけるバイアス
ISO 42001:2023	ISO	情報技術－人工知能－マネジメントシステム
ISO 23894:2024	ISO	情報技術－人工知能－リスク管理に関するガイダンス
ISO 12791:2024	ISO	情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理

前述の通り，この政策文書は，コンテンツの出所，情報の信頼性と真正性，そして透かしという3つの主要分野に焦点を当てています。より広範な標準化の状況を包括的に理解するために，本書を付随する専門技術の中枢報告書と併せてお読みいただくことをお勧めします。この中枢報告書では，マルチメディアの真正性確保の課題に対処する上でも重要な，資産識別子と権利宣言という2つの追加分野について詳細な分析を提供しています。また，これらの規格をどのように，どこで適用できるかについての実践的な推奨事項も提供しており，政策立案者と実装者の両方にとって貴重なガイダンスとなります。https://www.worldstandardscooperation.org/what-we-do/amas/

3.3 適合性評価：規格から保証へ

適合性評価とは，試験，検査，認証，監査などの方法を通じて，規格への適合性または規制への遵守を検証するプロセスです。政府および規制当局は，認証および適合性評価の結果に基づいて，製品が強制的な国家技術規制または任意規格の規定要件に準拠しているかどうかを判断します。ISO/IEC17000ファミリーなどの国際規格を基盤とする適合性評価は，技術規制および規格とともに，WTO/TBT協定で規定される3つの中核的な柱の1つです。製品，サービス，プロセス，請求システム，あるいは個人に関わるものであっても，プロセス全体を通して独立した保証が提供され，透明性が向上し，サプライチェーンの健全性が強化され，効率性，貿易円滑化，適合性検証が促進されます。

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?

2024年3月，WTO TBT委員会は，規制当局が適切かつ適切な適合性評価手続きの選択と設計を支援し，収束を図ることを目的として，非規範的な実務ガイドラインを公表しました。その基本原則は以下のとおりです。

- **非規範的 –** WTO加盟国に対し任意かつ拘束力を持ちません
- **中立的 –** WTO加盟国全体で，規制当局が適合性評価手続きに対して異なるアプローチを採用することを許容します
- **柔軟性 –** 適合性評価分野におけるアプローチとツールの革新を可能にすることを意図しています
- **補完的 –** 既存の業務やガイダンスに取って代わるものではなく，政府，規制当局，認定機関，その他による国家，地域，国際レベルの継続的な業務に貢献します

国際規格に関して，ガイドラインは次のように規定しています。「TBT協定第5条4項に基づき，加盟国は国際標準化機関が発行する関連ガイドまたは勧告を活用するものとする。例えば，加盟国はISO適合性評価委員会（CASCO）のツールボックスなどの適合性評価規格を活用することができる。

ただし，規制当局には，適合性評価のための国際規格，ガイド，または勧告の選択に制限はない。」

マルチメディアコンテンツ認証などの新興分野における適合性評価スキームを策定する際には，規制当局がTBT委員会の勧告と指針を考慮することを提案します。

例：EU AI法

同様のアプローチが採用されている分野として，提案中のEU人工知能法（EU AI法）に基づくEU適合性評価があります。適合性評価（CA）は，高リスクAIシステムがEU市場に投入される前，またはサービス提供が開始される前に，規制の要件を満たしていることを確認するための中心的なメカニズムです。以下の分野を対象にしています。

- リスク管理システム
- データガバナンス
- 専門文書
- 記録管理
- 透明性と情報提供
- 人による監視
- 正確性，堅牢性，サイバーセキュリティ

このフレームワークは，生成AIとマルチメディアコンテンツの認証にも同様のアプローチを開発すべきかどうかという疑問を提起します。

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.

規制当局と立法者がこの分野のガバナンスメカニズムを設計する際には，イノベーションの余地を維持しながら，信頼，説明責任，相互運用性を促進する上で，どの種類の適合性評価が最も適切かつ効果的な手段であるかを評価する必要があります。

WTO TBT委員会のガイダンスとEUモデルの両方を考慮することで，技術の進歩を阻害することなく，誤情報，偽情報，ディープフェイク，マルチメディアコンテンツの認証といった課題に対処するための強固な適合性評価スキームを開発するための確たる基盤が提供されます。

3.4 まとめ

マルチメディアの真正性，誤情報，生成AIに関連する複雑なリスクを管理するには，協調的な規格ベースのアプローチを採用する必要があります。国際規格は，規制の策定を導き，コンプライアンスを支援し，イノベーションを促進するための，信頼性が高く，実績があり，世界的に受け入れられているフレームワークを提供します。

これらの規格は，強固な適合性評価メカニズムと組み合わせることで，以下のことを実現します。

- 管轄区域全体での相互承認を促進します
- 相互運用性と信頼性を実現します
- 重複とリソースの非効率性を削減します
- 消費者を保護し，公共政策の目的を遵守します

結局のところ，前述のPDRフレームワークの有効性は，それを支える規格と保証システムの有効性に左右されます。本書の後半では，これらのツールを様々な分野やステークホルダーグループに実践的に適用し，AI駆動型マルチメディアコンテンツの真正性，倫理性，信頼性を確保する方法について考察します。

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.

第4章

技術的ソリューションとガイダンス

4.1　誤情報対策におけるコンテンツの出所の役割

コンテンツの出所を保証する技術的ソリューションは，マルチメディアの真正性を検証する上で不可欠です。これらのツールは，メディアの起源，履歴，変遷に関する情報を経時的に記録し，透明性の高いデジタル証跡を作成することで，誤情報の拡散を防ぎ，社会の信頼を再構築することを目的としています。

コンテンツの出所と真正性に関する連合（C2PA）は，テクノロジー企業とメディア組織の連合体であり，コンテンツ認証情報として知られるデジタルコンテンツの来歴に関するオープンな技術標準の策定を使命としています。300社を超える会員で構成されるこの連合は，Adobe，Amazon，BBC，Google，Intel，Meta，Microsoft，OpenAI，Publicis Groupe，Sony，Truepicからなる運営委員会によって主導されています。EUの2022年版「偽情報に関する強化行動規範」と，AIパートナーシップによる「合成メディアにおける責任ある実践のフレームワーク」は，このプロジェクトをデジタルコンテンツの透明性と真正性を高めるための有効な手段と位置付けています。

この分野におけるもう一つの先進的な事例として，コンテンツ真正性イニシアチブ（CAI）が挙げられます。CAIは，コンテンツ認証情報の推進を通じて，出所に関する技術，政策，教育の課題に取り組んでいます。CAIは，テクノロジー企業，学術機関，メディア組織，NGOからなる幅広いアライアンスで構成され，出所規格の世界的な導入を促進するために協力しています。

出所規格の分野はまだ成熟段階ですが，CAI，C2PA，ISO，ITU，IECによる共同イニシアチブは急速に進展しています。これらのイニシアチブには，コンテンツの出所を追跡し，改ざんを検出し，デジタルメディアの信頼性を確立するためのツールや方法論が含まれています。

コンテンツ認証情報は，ISO規格化に向けて加速しています。ISO/CD 22144「情報の真正性 – コンテンツ認証情報」は，コンテンツの出所と認証に関する国際規格として正式に認められる可能性があり，近い将来，正式に認められる可能性があります。この規格は，メディアの出所を記述する暗号署名付きメタデータを規定しており，ソフトウェアからのエクスポート時，あるいはハードウェアでの作成時にメディアコンテンツに添付することができます。耐久性のあるコンテンツ認証情報（DCR）を使用することで，メディアに電子透かしを追加し，堅牢なメディアフィンガープリント照合システムを実装することで，コンテンツ認証情報の検索のための2つの追加保存レイヤーを組み込むことができます。

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.

4.2 補完的な取り組み

WITNESS

WITNESSは30年以上にわたり，人々が人権擁護のためにビデオやテクノロジーを活用し，信頼できる情報を共有できるよう支援してきました。最近，AI生成メディア，特に視聴者を誤解させる可能性のあるハイパーリアリスティックなシミュレーションの作成がもたらすリスクについて懸念を表明しています。

WITNESSは規格にのみ焦点を当てているのではなく，透かし，ラベリング，検証システムの倫理的利用を導くCAIやC2PAといったフレームワークの導入も支援しています。これにより，真正性と人権およびアクセシビリティへの配慮のバランスが保たれます。

MAVEN

MAVENコンソーシアムは，コンテンツ認証とマルチメディア分析ツールを「検索と検証」機能に重点を置いた統合プラットフォームに組み入れることを目指しました。しかし，この取り組みは，その強力な基盤目標にもかかわらず，より広く認知された代替手段との競合により，普及が限定的となっている可能性があります。

JPEG Trust

JPEGは，新しい国際規格「ISO/IEC 21617-1:2025 情報技術 - JPEG Trust」の開発を開始しました。この規格は3部構成で，メディアにおける信頼を確立するためのフレームワークを規定しており，出所，真正性，完全性，著作権，資産および利害関係者の特定といった側面を含んでいます。

IEEE AI倫理グローバルイニシアチブ

自律型およびインテリジェントシステムの倫理に関するIEEEグローバルイニシアチブ2.0は，グローバル志向，学際的連携，包摂性，実践倫理という4つの柱を重視しています。このイニシアチブは，規格，ツールキット，認証ツールを推進し，IEEE 7000シリーズの採用を奨励しています。

草の根および人権擁護活動

Guardian ProjectやOpenArchiveなどの組織は，ObscuraCam，InformaCam，ProofModeなどのモバイルアプリを活用し，暗号的に検証可能な写真，動画，音声のキャプチャをサポートしており，ジャーナリズム，アクティビズム，アーカイブのための記録を強化しています。

4.3 新たな共通点

これらの多様な取り組みの中で，デジタル署名，出所追跡メカニズム，標準化されたメタデータモデルといった共通機能が出現しています。これらの要素は，特に公開鍵基盤（PKI）の利用拡大に伴い，マルチメディアの真正性の運用と規制に不可欠であるとますます認識されています。ストリーミングプラットフォーム，コンテンツ制作者，ソーシャルメディアサービスが不正行為に対処し，信頼性を確保しようとする中で，これらの機能の統合はますます重要になっています。

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?

第5章

規制の策定と適合性の向上を支援する：政策立案者と技術提供者向けチェックリスト

マルチメディアの真正性に対する信頼を構築するために，規制当局と技術提供者が規制や執行フレームワークを策定する際，あるいは技術ソリューションを開発する際に活用できる以下のチェックリストを提供しています。このチェックリストは，期待値の整合，ギャップの特定，責任あるイノベーションの促進，そして適合性の向上に役立ちます。

領域	規制当局向け設問	技術提供者向け設問
対象範囲	対象となるコンテンツの種類は？	どのような種類のコンテンツを提供していますか？
	どの省庁や機関の関与が必要ですか？	ツールは，業界固有の規制に対応するようにカスタマイズされていますか？
規制要件と執行	措置は任意ですか，それとも強制ですか？罰則はどのようなものですか？	遵守すべき規格や適合性評価制度はありますか？それらに対応する準備はできていますか？
	執行能力はどの程度ですか？規制機関はありますか？	関連する規制当局を認識していますか？
規格の支援	どのような自主的な国際規格が，あなたのアプローチを強化したり，目標達成を支援したりできますか？	安全で安心な開発を支援するための関連規格や適合性評価はありますか？
	システムと関連機関は，どのように目標達成に役立ちますか？	QIシステムは，どのようにソリューションの信頼性を高めることができますか？
技術的選択肢	コンテンツライフサイクルの各段階でどのようなツールが利用可能ですか？それらは規格に基づいていますか？	ツールは，法規制や技術の発展に合わせて進化していますか？
	それぞれの利点と限界は何ですか？	ツールの強みと限界を明確に伝えていますか？

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.

以下は，規制当局，政策立案者，技術提供者が，選挙運動，自然災害，危機管理などの状況で使用できる追加のチェックリストと，使用順序です。

アクション	文書	説明	規制当局，立法者，政策立案者向け	技術提供者と実装者向け
1: 全体像を把握するために，まずこのチェックリストから始めます。	初期チェックリスト	出発点として使用	このチェックリストは，すべての関係者がそれぞれのニーズについて意見を述べる機会を確保するために使用します。	政府，立法者，または規制当局が関係者に期待する回答を明確にするために使用します。
2: 主要なリスクを特定するためのPDRを作成します。発生しているシナリオに基づいて使用します。 例えば，選挙運動の場合は，そのためのPDRを作成します。 自然災害のような事態の場合は，それぞれのリスクに特化した新しいPDRを作成します。	i誤情報，偽情報，ソーシャルメディアに関するPDR	3つの柱を詳細に説明したPDR	ソーシャルメディアからの誤情報，偽情報のリスクを保護，検知，対応するための補助として活用します。 規制当局，立法府，または政府の組織的活動から得られる情報を保護し，国民が受け取る情報の信頼性を継続的に確保するために活用できます。	企業やソリューションプロバイダーに影響を与えるソーシャルメディアからの誤情報，偽情報のリスクを保護，検知，対応するための補助として活用します。
3: PDRの出力結果に応じて，最大のリスクが何であるかが明らかになります。 マトリックスを使用して，必要な保証レベルまたは信頼性を提供する準拠すべき規格を選択します。	MMCAマトリックス	これは色分けされたマトリックスで，既存の規格，ガイダンス，規制をリストアップしており，様々な組み合わせを用いることで，様々なトピックにおいて異なるレベルの保証を提供できます。 組み込まれるセットが大きいほど，保証レベルが高くなります。	規制当局や政策立案者は，適合性評価スキームに参照または組み込むことができる，既存および新興の技術や規格を検討するための出発点として使用できます。	概説されている技術を用いて自主規制を検討し，どの程度の保証レベルを構築できるかを検討したい組織にとって有益です。

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.

アクション	文書	説明	規制当局，立法者，政策立案者向け	技術提供者と実装者向け
4: あらゆるシナリオにおいて，このチェックリストを使用して，適切な質問が行われ，チェックが実施されていることを確認する。	マルチメディアコンテンツ認証チェックリスト	様々な用途におけるマルチメディアコンテンツの真正性を判断するのに役立つ質問をまとめたスプレッドシート	規制当局や政府機関がコンテンツを検証するために使用できます。政策立案者や立法者は，監査人や適合性評価機関に対し，テクノロジーソリューションプロバイダーのチェック内容を確認するよう促すためにも使用できます。	報道機関やその他のメディアプラットフォームなどの組織が，送信または消費するコンテンツの真正性を検証するのに役立ちます。
5: 上記の4と並行して使用することも，時間的な制約により迅速な回答が必要な場合は，上記の4の代わりに使用します。	一般的なチェックリスト	検討すべき設問をまとめた簡潔なチェックリストで，より具体的な透かしソリューションのチェックリストが含まれています。	規制当局，立法者，政府がデジタルコンテンツの真正性を評価する方法を検討する際に使用できます。	迅速な真正性チェックを実施したい組織やメディアプラットフォームに役立ちます。
6: 透かしに関する設問への具体的な回答が必要な関係者がいつでも使用します。	透かしチェックリスト	透かしをより詳細に検討する特定のチェックリスト	規制当局や政策立案者は，透かしに大きく依存するソリューションの適合性評価スキームにおいて，参照または組み込むことができる既存および新興の技術や規格を検討するための出発点として使用できます。 政策立案者や立法者は，監査人や適合性評価機関に対し，技術ソリューション提供者のチェック内容を確認するよう促すために使用できます。	透かしソリューションを開発する技術ソリューション提供者は，高品質な製品を確保し，リスクを最小限に抑えるために，どのような既存および新興の技術や規格に従うことができるかを検討するために使用できます。 透かしの使用を希望する，または透かしソリューションに大きく依存している組織に役立ちます。

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.

アクション	文書	説明	規制当局，立法者，政策立案者向け	技術提供者と実装者向け
7: 利用可能なツールに関する意思決定，または保証が必要な段階にあるツールに関する意思決定を支援するために，他のチェックリストと並行して使用します。	マルチメディアコンテンツ認証ガイダンスのための最新および新興の技術	本書を補足する資料として，現在認証に使用されている技術について，より技術的な概要を説明しています。	規制当局，政策立案者，監査機関，標準化団体が，適合性評価スキームに参照または組み込むことができる既存および新興の技術や規格を検討するための出発点として活用できます。	ここで概説した技術を用いて自主規制を検討し，保証技術によって信頼性を構築したいと考えているソリューション提供者や組織にとって有用な手段です。

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.

第6章

推奨事項

以下の推奨事項は，国際および国内の政策立案者，規制当局，メディアと技術分野，規格開発団体（SDO）を対象としています。これらの推奨事項は，マルチメディアコンテンツの真正性を強化し，世界的な信頼を築くために，迅速に実践可能です。

政策立案者および規制当局向けの事項は以下のとおりです。

- 第4章に記載されているチェックリストを検討してください。
- 特に規制の整合性を促進するため，多国間フォーラムを通じて，標準化および整合性に関する取り組みに参加し，連携してください。
- 規制サンドボックスを開発・導入する際には，管理された環境で新しい技術，政策アプローチ，コンプライアンスモデルを試験するための国際規格を考慮します。
- コンテンツの真正性に関する課題への対応策を策定するため，国際的に認められた規格に基づくPDRフレームワークを採用します。
- AI生成コンテンツがユーザーの権利を尊重し，差別的な結果を回避できるよう，データプライバシーとバイアスに関する規制を考慮します。
- AIリスク，誤情報，偽情報，ディープフェイクに関する要件を組み込んだ，マルチメディアコンテンツに特化した適合性評価フレームワークの開発を支援・奨励します。
- マルチメディアコンテンツの認証に関する適合性評価および／または認証スキームについて，関連する試験を含む保証を提供する国際規格に基づくものを検討します。

技術開発者および提供者に対して，政策立案者は以下の事項を検討するよう要請できます。

- コンテンツの真正性に関する課題への対応策を体系化するために，国際的に認められた規格に基づくPDRフレームワークを採用すること。
- 規制要件を満たし，将来を見据えたイノベーション・パイプラインを構築するために，国際規格およびベストプラクティスに準拠し，監視すること。
- 組織内に規格のリエゾン担当者または推進者を任命し，最新情報の追跡，コンプライアンスの確保，新たな要件の導入支援を行うこと。
- 安全なマルチメディア認証とコンテンツの整合性を実現するために，PKIなどの強力な暗号化プロトコルの統合を検討すること。
- 安全なタイムスタンプ，改ざん防止ハッシュ，デジタル署名を活用し，ユーザーのプライバシーを保護しながらコンテンツの真正性を検証すること。

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.

第7章

結論

本政策文書では，AI生成コンテンツとマルチメディア操作がもたらす喫緊の課題，特に誤情報，偽情報，ディープフェイクといったコンテキストにおいて，その課題を検証しました。堅牢な国際規格と適合性評価フレームワークに支えられた，協調的な国際的対応の緊急性を強調しています。

本書は，透かし，コンテンツの来歴，真正性という3つの主要分野に焦点を当て，PDRフレームワークなどのツールを活用することで，規制当局，業界，規格団体が協力してリスクに対処し，イノベーションを維持するための具体的な手順を概説しています。

重要なのは，本書で提示された推奨事項とチェックリストが，政策と実践のギャップを埋め，生成AIと関連技術を安全，倫理的，かつ包摂的に活用できるようにすることです。これらの対策が先進国と発展途上国の両方で一貫して実施されれば，マルチメディアコンテンツの信頼性，検証可能性，そして公共の利益との整合性を確保するのに役立ちます。

結論として，実用的なガイダンスと認証に裏付けられた国際規格の効果的かつ調和のとれた活用は，安全で透明性が高く，イノベーションに配慮したデジタル情報エコシステムへの確かな道筋を提供します。

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.

附属書1

誤情報，偽情報，および悪意の情報の種類は多岐にわたります。例えば，以下のようなものがあります。

捏造されたコンテンツ	通常，100%虚偽であり，欺瞞と危害を加えることを目的として作成されています。[15] 本物のコンテンツと捏造されたコンテンツを区別することは非常に困難です。捏造されたコンテンツを宣伝するために使用される高度なディープフェイクに遭遇すると，市民が受け取るメッセージへの信頼に深刻な影響を与える可能性があります。
操作されたコンテンツ	歪曲された本物の情報または画像。この種のコンテンツは，画像を加工したり，センセーショナルな見出しやクリックベイトを使用したりすることで，本物のコンテンツを操作していることがよくあります。
なりすましコンテンツ	本物の情報源を偽装したもので，多くの場合，既存の通信社や評判の良い報道機関のブランドを使用しています。この形態の偽情報は，人々が特定の組織，ブランド，あるいは個人に対して抱く信頼を悪用します。攻撃者は，有名ブランドを装ったフィッシングやスミッシングのメッセージを用いて，受信者が正当なコンテンツを受け取っているという印象を与えようとします。
l誤解を招くコンテンツ	誤解を招く情報は，見出しで記事を再構成することで作成されます。これは通常，より広範な論点を裏付けるために引用の断片を用い，多くの場合，自身の立場に沿う形で統計を引用します。あるいは，主張を弱める可能性があるため，意図的に特定の記事を取り上げないことも考えられます。主張を展開する際には，誰もが自分の主張全体を裏付ける内容を持ち出そうとします。
偽のコンテキスト	事実に基づいた正確なコンテンツに，虚偽のコンテキスト情報が組み合わさったもの。例えば，記事の見出しが内容を反映していないなどです。つまり，本来のコンテンツが再構成されているのです。虚偽のコンテキスト画像は，ローテクではありますが，それでもなお強力な誤情報や偽情報の形態となります。
風刺とパロディ	ユーモラスだが虚偽の情報を真実として偽装したもの。害を及ぼす意図はないものの，読者は騙される可能性があります。かつては芸術の一形態として扱われていたものが，今では意図的に噂や陰謀を広めるために盛んに利用されています。加害者は，真剣に，あるいは文字通りに扱うべきではないことをしているだけだと主張するため，取り締まりは困難です。この種の誤情報や偽情報の危険性は，それが再共有される方法と速度にあります。その過程で，情報はしばしば再構成され，より広範な読者が元の発信者とのつながりを失い，風刺として理解できなくなります。
誤ったコネクション	扇情的な見出しやクリックベイトのような見出し，画像，キャプションが記事の内容をサポートしていない場合。表面的には，この種のコンテンツは単に不快感を与えるものと見なされるかもしれませんが，効果的に使用されると，メディアへの信頼を損ない，分断を助長する可能性があります。サイトへのダイレクトトラフィックの必要性が高まるにつれて，信頼と報道機関の関係は弱まる可能性があります。
スポンサーコンテンツ	編集コンテンツを装った広告またはPR。これは，誤情報や偽情報の悪用による影響は小さいように見えるかもしれませんが，真の報道機関にとっては利益相反となる可能性があります。消費者が，コンテンツが広告であることを容易に識別できない場合，不適切なラベル表示によって意図的に誤解を招いていると主張できます。

[15] この種類は，2023年11月にソーシャルメディアで広く拡散されたロンドン市長サディク・カーンのディープフェイク音声クリップのような虚偽のコンテンツを使用します。演者は，同日に行われる追悼週末の式典よりも親パレスチナデモを優先するよう呼びかける市長の声を模倣しました。

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.

プロパガンダ	態度，価値観，知識を管理するために使用されるコンテンツ。プロパガンダは常に，プロパガンダの意図を促進する反応を得るために，認識を形成し，認知を操作し，行動を誘導するための体系的な試みとして使用されてきました。伝統的に，プロパガンダは互いに積み重なる複雑なメッセージで構成されていました。現在，プロパガンダはAI，ボット，トロール，フェイクニュースサイトなどを利用して，メッセージを広く迅速に拡散しています。その方法として，その効果はより直接的かつ即時的です。
エラー	既存の報道機関が報道において犯す間違い。ニュースが存在する限り，エラーはニュースに存在してきました。誤情報と偽情報が報道機関にもたらす問題は，やはりスピードに関係しています。速報をいち早く伝えようとする努力によって，信憑性の確認にかかる時間が最小限に抑えられてしまいます。その結果，報道機関は，正当な記者とされる人物から送られてきたAI生成コンテンツやディープフェイクコンテンツに翻弄されることになります。

附属書2

この問題の性質は，以下を含む多くの利害関係者に影響を与えます。

有権者：選挙期間中，マルチメディアによる真正性のないAI強化された誤情報が意図的に拡散されることは，有権者にますます大きな影響を与えています。こうした利用は，有権者を意図的に混乱させ，偏見を生み，民主主義国家において選挙結果を歪めることになります。さらに広く言えば，こうした行為は，権力機関や従来型メディアに対する国民の信頼を損ない，疑念や幻滅感を生みます。

消費者（継続的な影響）：予測分析や自動広告ターゲティングなどのAIツールを消費者向けシナリオで活用することで，消費者と企業の両方にメリットをもたらすことができます。これらのツールは，かつてないほどの効率性と顧客インサイトをもたらし，パーソナライズされた顧客体験を提供することができます。残念ながら，これはマイナスの影響も生み出します。消費者はAI疲れに陥り，AIを活用したコンテンツの氾濫によって不誠実さを感じ，真の人間的な繋がりを求めるようになります。コンテンツが認証されていない場合，この傾向は悪化し，消費者が詐欺の被害者になることもあります。

悪意のある人物が認証されていないマルチメディアを用いて詐欺や情報操作を目的とした偽コンテンツを作成した場合，個人は金銭的損失や人身的危害を被る可能性があります。さらに，認証されていないコンテンツは，ユーザーの追跡，個人情報の窃盗，マルウェアの拡散にも利用される可能性があります。個人の購買行動を追跡するAI生成ポップアップは，その性質上，購入への衝動を煽る強制力となります。ポップアップが悪意のある攻撃を受けると，即座に脅迫メッセージが表示される可能性があります。また，消費者は，製品の真の支持者によるものとされるマルチメディアコンテンツを提供するサイトに誘導される可能性もあります。

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.

攻撃者はAIを駆使して，有名人に似た画像を生成したり，本物と見分けがつかないほど正確に声を模倣した音声クリップを作成したりしています。これは，例えば精神疾患に関連する理由で，短期間で体重を減らそうとしたり，不安やうつ症状を緩和しようとしたりするような，社会で最も弱い立場の人々に影響を与えることがよくあります。同様に，がん治療分野のいわゆる医療専門家による科学的に聞こえる記事や動画を利用した誤情報や偽情報が，長年デジタルプラットフォーム上で蔓延しています。ハーブや代替療法が化学療法の必要性をなくしたり，症状を緩和したりできるという主張は一般的であり，偽りの希望を与えます。これらの動画に，著名な権威者や著名人が製品を推奨しているように見えるディープフェイクが組み込まれている場合，説得力は増します。

投資詐欺が増加しており，最近ではFacebookで，著名な英国のファイナンシャルアドバイザー，マーティン・ルイス氏とIT界の大富豪イーロン・マスク氏のディープフェイク動画が使用され，存在しないビットコイン投資スキームが宣伝されました。また，ITVの政治アナリスト兼コメンテーター，ロバート・ペストン氏も仮想通貨投資を推奨している姿が見られました。

概念的に言えば，マルチメディアの信憑性に左右されないこの種の誤情報や偽情報は，人間の感情を操作することを前提としています。消費者は，ブランドが自分と価値観が似ている権威者や著名人によって支持されているという考えに惑わされ，これに騙されやすいのです。

この問題への対応として，FacebookとInstagramを所有するMetaは，広告で著名人を不正に利用する詐欺師を取り締まるため，顔認識技術を導入する予定です。

政治家（一貫して）：多くの人にとって，政治家の信憑性は，政治候補者の評価と投票行動の基盤となるものです。私たちの判断は，テレビニュースの視聴率，ソーシャルメディアで読んだり見たりした政治記事，そして候補者のプロフィールに基づいています。ほとんどの人はそれぞれ独自の政治的態度や考えを持っていますが，メディアから得る具体的な印象も大きく影響します。そして，それが，政治家が対立候補よりも信憑性が高い，あるいは低いという私たちの認識に影響を与えます。

政治家と直接会話する機会を持つ人はほとんどいません。その結果，政治家の信憑性，信頼性，誠実さの評価は，メディア情報によって形成される印象に依存しています。政治家へのテレビインタビューが始まった当初は，人々は画面上の人物像を個人的に知っているという認識を通して，政治家を評価する力があると感じていました。今日では，ソーシャルメディアとポピュリズムは，直接的で即時的なコミュニケーションスタイルによって，相互に強化し合う関係性を強化しています。候補者のソーシャルメディア上での自己紹介は強力なツールであり，政治家はこれを用いて，従来型メディアや制度化されたメディアの制約を受けずに，より個人的な方法で国民に直接語りかけているような錯覚を与えることができます。

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/

これは，マルチメディアの信憑性の欠如のリスクや，デジタルプラットフォームにおけるフェイクニュースの拡散が民主的な統治の質を損なうという懸念を考慮するまでは，前向きな変化であるはずです。これは，政治家が自らに有利にも不利にも利用できる要因です。

アーティスト（経済的）：もう一つの重要な懸念は，芸術界におけるAI作成コンテンツの大規模な拡散です。これは，既に深刻なデジタル誤情報の問題を悪化させています。AIツールは，詐欺師，ペテン師，犯罪者に，記事，音声，画像，写真，動画，歌，芸術作品など，人工的なコンテンツや虚偽の情報を作成する強力かつ効果的な手段を提供します。元のクリエイターの肖像やスタイルを模倣して人工的に作成された場合，偽物や虚偽であると見分けるのは困難です。 AIツールが悪意ある目的で意図的に悪用されるだけでなく，AIが作成したコンテンツは純粋に人間が作成したコンテンツよりもはるかに高速に作成できるため，信憑性は急速に低下します。

一般市民：COVID-19パンデミックの発生は，フェイクニュースの波を引き起こしました。ウイルスの治療法に関する誤ったアドバイスが人命を危険にさらすなど，誤情報と偽情報が世界中に蔓延しました。トランプ大統領が記者会見で，COVID-19患者に消毒剤を注射するというアイデアについて「興味深いですね」と述べ，「消毒剤が1分で患者を死亡させるのが分かります。1分です！」と発言したことや，5G基地局がCOVID-19と何らかの関連があるという主張が当時広く報道されました。これは，良くても混乱を招き，最悪の場合，事態を収拾しようとする当局への不信感を招きました。パニックと孤立の時代に，多くのコメンテーターが虚偽の主張を反駁しているにもかかわらず，市民はこうした話に非常に影響を受けました。誤情報の共有は人々の精神的健康に影響を与え，ひいては健康全般にも影響を及ぼす可能性があります。

ソーシャルメディアは，有害な成分に関する主張から，政府がワクチンを人口制御に利用したという陰謀論に至るまで，ワクチンの安全性に対する個人の認識に大きな役割を果たしました。人の意思決定に不当に影響を与えることは，最終的には壊滅的な結果をもたらす可能性があります。

個人やグループがソーシャルメディアを利用して，様々な目的を持つ誤情報や偽情報を拡散する容易さと速さから，世界保健機関（WHO）は「インフォデミック」という用語を，また他の団体は「ディスインフォデミック」という用語を用いています。特に，アフリカ系住民は免疫があるといった，根底に人種差別的または外国人排斥的なニュアンスを持つ偽情報に対抗するために，神話を打ち破るキャンペーンが必要となりました。

若者：2021年に英国セイファーインターネットセンターが実施した調査では，「若者の半数が日常的にオンラインで誤解を招くコンテンツに遭遇している」ことが調査されています。さらに，この調査では，「若者の48%が毎日誤解を招くコンテンツを目にしており，10人に1人以上が1日に6回以上目にしている。その結果，しばしば苛立ち，動揺，悲しみ，怒り，攻撃されていると感じたり，恐怖を感じたりしている」ことも明らかになっています。[16]

この状況は，依存者が直面するリスクや危害を合理化できるものの，依存から抜け出すことができない依存症に似ています。

[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.

若者は，大人に比べてメディアリテラシーが低い傾向があり，ニュースについて批判的に考えたり，マルチメディアの信憑性に疑問を呈したりするだけの十分な認識を持つことが難しい傾向があります。オンラインプラットフォームや提供されるコンテンツに過度にさらされることで，若者は不安を助長する状況に陥りやすくなり，自尊心が低下し，過激な意見が植え付けられ（それが若者の信念や行動に深刻な影響を及ぼす），虚偽の記憶が植え付けられ，悲惨な見通しにつながる可能性があります。

有害コンテンツは拡散しやすく，ソーシャルメディアが生み出す他の現象，例えば他の若者の生活に対する理想化や非現実的な見方などとの相互関係によって，特に危険となります。コンテンツに関する統制とガバナンスの欠如により，誤情報や偽情報は，真正性管理の欠如によって生じる空白を悪用します。単純な画像でさえフィルタリングによって操作され，完璧な容貌や体格をリアルに描写するようになります。ボディイメージの問題に敏感で，目にする完璧な画像に太刀打ちできず，周囲に適合しなければならない，あるいは「合わせなければならない」と感じている若者は，自傷行為，拒食症，過食症，あるいは自殺に関連する内容を助長する有害なコンテンツの影響を受けやすくなります。

サイバーいじめやチャイルド・グルーミングは，加害者によって新たな技術革新が巧妙に利用され，ますます巧妙に促進されています。

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.

附属書3

ディープフェイク：カテゴリーと脅威ベクトル

ディープフェイクは，生成AI（GAN，VAE，トランスフォーマーなど）を用いて作成された，操作された，あるいは完全に生成された合成メディアです。メディアの種類と意図によって分類されます。

種類	方法	脅威
オーディオディープフェイク	音声複製：少量の音声サンプルを用いて個人の声を模倣する（例：CEOのなりすまし） 合成音声生成：偽のスピーチや会話を作成する	ソーシャルエンジニアリング（CEO詐欺），偽情報，電話詐欺
画像ディープフェイク	顔スワッピング：動画や画像内で人物の顔を別の人物に置き換える リップシンク：新しい音声に合わせて唇の動きを変える 表情操作：感情や行動を変える	偽情報キャンペーン，風評被害，恐喝
ビデオディープフェイク	全身蘇生：体のジェスチャーや動きを完全に生成する ポーズトランスファー：ある人物のポーズを別の人物の体にマッピングする	脅威：政治操作，虚偽の自白，スパイ活動
テキストディープフェイク	合成ニュース／ブログ：偽の記事や文書を生成する 偽のチャットボット／メール：テキストベースの会話におけるなりすまし（例：フィッシング）	フェイクニュースの拡散，自動トローリング，フィッシング
画像ディープフェイク	AI生成ペルソナ：詐欺や監視回避に使用される実在しない顔 画像間変換：画像のビジュアルスタイル／コンテンツの変更（例：オブジェクトの削除，背景の変更）	ソックパペット，詐欺，偽ID，誤情報

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.

生成AIを活用したサイバー攻撃

生成AIは，従来型および新型のサイバー攻撃に新たなベクトルを提供します。

種類	方法	脅威
フィッシングとソーシャルエンジニアリング	大規模なスピアフィッシング：AI生成のカスタマイズされたフィッシングメール ボイスフィッシング（ヴィッシング）：標的を欺くために複製された音声 ディープフェイク動画フィッシング：幹部を装った偽のZoom/Teams通話	認証情報の盗難，不正アクセス，BEC（ビジネスメール詐欺）
マルウェアとエクスプロイトの生成	マルウェア用コード生成：AIがポリモーフィック型マルウェアまたはシェルコードを生成する 難読化と回避：GPTのようなモデルが，既知のマルウェアの検出不能な亜種を作成する	エンドポイントの侵害，データの窃盗
誤情報と偽情報による攻撃	AI生成フェイクニュース：大規模なナラティブ操作 合成インフルエンサー：合成ペルソナを持つボットがプロパガンダを拡散する	選挙干渉，経済操作，風評被害
なりすましとID詐欺	合成ID作成：GANを用いて偽造IDやIDポートフォリオ全体を生成する 音声/顔IDのなりすまし：合成入力を用いて生体認証システムを回避する	銀行詐欺，KYC（顧客確認）の回避，監視の回避
データポイズニングとモデル攻撃	学習データの操作：AIモデルの学習に悪意のあるデータを挿入する プロンプトインジェクション攻撃：細工された入力を用いてLLMを悪用する	モデルの劣化，誤分類，不正な行動
Cコンテンツフラッディングと信頼へのDDoS攻撃	情報過多：生成AIはプラットフォームに偽コンテンツ（レビュー，苦情，ニュースなど）を氾濫させる	過剰なモデレーションシステムにより，情報源の信頼性が損なわれる

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.

ハイブリッド型および新たな脅威の種類

マルチモーダルディープフェイク:

音声，動画，テキストを組み合わせて，より説得力のある欺瞞を行う。

自律型AI攻撃エージェント:

サイバー攻撃を自律的に計画・実行するLLM。

敵対的サンプル生成:

AI検出／分類システムを欺くためにわずかに改変された画像／ビデオ。

セクストーションやリベンジポルノのための合成メディア:

脅迫に使用される偽の親密な画像。

防御上の考慮事項

以下は，役立つ可能性のある防御アプローチのほんの一部です。チェックリストで詳細に説明しています。

検出ツール:

透かし，フィンガープリンティング，敵対的検出ツール，フォレンジックツール

検証プロトコル:

暗号署名，多要素検証

政策とガバナンス:

AI監査，法的フレームワーク，倫理規格

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.

誤情報 偽情報 ソーシャルメディア

種類	意図	防止 すべて	保護 政府関連	検出 すべて	対応
偽のペルソナやウェブサイトの作成	偽のペルソナやウェブサイトのネットワークを構築し，ターゲットオーディエンスにおけるメッセージの信憑性を高めることで，偽情報を拡散することを目的としています。典型的には，偽の学術専門家，ジャーナリスト，シンクタンク，学術機関などが挙げられます。偽の専門家ネットワークは，偽の資格情報を用いてコンテンツの信憑性を高めます。	オーディエンスを公式ウェブサイトや信頼できる情報源に誘導するようにしてください。ウェブサイトでは，人々が信頼できる情報源として頼れる，明確で簡潔かつ最新の情報を提供するようにしてください。オンライン情報は最新の状態に保ってください。 組織，主要な代表者，広報担当者のすべてのソーシャルメディアアカウントを検証してください。記事，論文，その他のリソースを共有する前に，その情報源を確認してください。	政府機関は，ウェブサイトを.govトップレベルドメインに移行し，政府機関のみが利用できる.govドメインを使用して，ウェブサイトが本物で安全であることを国民に周知する必要があります。	セマンティックチェッカーを使用して定期的にスキャンを実施してください。	管轄区域の法律または規制がある場合は，それらに基づいてコンテンツの削除を強制してください。
合成メディアとディープフェイクの作成	攻撃者は，これを利用して，誰かが実際には行っていないことを行ったり，言っていないことを言ったりしているように見せかけます。 偽情報キャンペーンの一環として，合成メディア技術を，悪意を持って使用し，虚偽の情報を共有したり，オーディエンスを操作したりします。	個人情報が合成メディアコンテンツの生成に利用される可能性があることを周知するための啓発キャンペーンを実施してください。個人アカウントと業務アカウントの両方で，適切なサイバー衛生対策を実施してください。 リバースイメージ検索などの公開ツールを活用して，メディアコンテンツのソースを確認してください。合成メディアを含むコンテンツを共有または作成する場合は，たとえ無害な用途であっても，免責事項を追加し，一般人の意識を高めてください。 ディープフェイク動画や音声クリップへのインシデント対応計画を策定してください。		組織やメッセージに影響を与える合成メディアを迅速に特定し，公式チャンネルで反論し，可能であれば証拠を提示してください。 ソーシャルメディアプラットフォームに適用可能なコンテンツ真正性検証ツールを使用してください。	管轄区域の法律や規制がある場合は，それらに基づいてコンテンツの削除を強制してください。 被害者にディープフェイクについて注意喚起するための情報キャンペーンを実施してください。 一般人がディープフェイクや合成コンテンツを識別できるように，啓発プログラムを実施してください。

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	

種類	意図	防止 すべて	保護 政府関連	検出 すべて	対応
陰謀論（新たな陰謀論の考案または既存の陰謀論の増幅）	陰謀論の視点に沿った偽情報のナラティブを作成することで，ターゲットオーディエンスの共感を呼ぶ陰謀論を活用します。特定の表現を複数回繰り返して使用することで 複数のナラティブを使用し，特定の表現を繰り返すことで，ターゲットオーディエンスのナラティブへの親しみを高め，信憑性を高めます。過激化を促進するために	ウェブサイトを常に最新の状態に保ち，明確で正確な情報を掲載してください。 オンラインとオフラインの両方のチャネルを構築し，同僚やパートナーと情報を共有し，信頼できる情報を広めるネットワークとして協力してください。 陰謀論の仕組みや，よく見かける画像や比喩表現について，視聴者に啓発キャンペーンを実施してください。	「偽情報」または「噂のコントロール」ページを作成し，維持管理することで，所属部署に関する偽ニュースや噂を即座に見破ることができます。	組織に影響を与える合成メディアがないか，定期的にサイトをスキャンしてください。 他者と協力し，敵対者や脅威アクターに関する情報を共有してください。	
情報フラッディングとアストロターフィング	アストロターフィングとは，同じナラティブを繰り返し，草の根レベルで広く支持または反対されているという印象を与えることで，メッセージへの信憑性を高める行為です。その真の出所は通常隠蔽されています。 偽アカウントや自動アカウントを用いて，ソーシャルメディアへの投稿をフラッディングやファイアホース攻撃でスパムし，反対意見を封じ込める行為です。多くの場合，偽アカウントや自動アカウントが多数使用されます。	信頼できるコミュニケーターのネットワークを地域内に構築し，信頼性が高く正確な情報を促進してください。組織がアストロターフィングやフラッディング攻撃の標的になった場合に備えて，複数のチャネルを活用して情報を共有してください。 オンラインとオフラインの両方のフォーラムを通じて，有権者からの議論，討論，フィードバックを促進してください。	信頼できるコミュニケーターのネットワークを構築するために，職員を活用してください。 他の政府系メディアチャネルを活用して，意識向上を図り，偽情報に対抗してください。 信頼できるインフルエンサーがホストする広報ナラティブ／広告を通じて，有益なメッセージを広めるために，人気の非政府系フォーラムを活用してください。	真正性チェックを実施してください。アカウントが不正である疑いがある場合 1）アカウント作成日などの詳細情報を確認 2）プロフィール写真と自己紹介 3）フォローしている他のサイトやアカウントを調査 4）投稿アクティビティを調査 5）ボットや荒らしの疑いのあるアカウントによる投稿ではないか確認	

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.

種類	意図	防止 すべて	保護 政府関連	検出 すべて	対応
	多くの場合，新しい政策や取り組みに関する議論など，正当な議論を制限または阻止し，人々がオンライン空間に参加することを阻止することを目的としています。情報操作者は，フラッディングを繰り返し行うことで，標的の感受性を蝕みます。真実は何もないという印象を与えることを目的としています。				
他のプラットフォーム／小規模コミュニティプラットフォームの操作	プラットフォームやコンテンツのモデレーションポリシーが厳しくなく，不正コンテンツの検出・削除のための管理体制が弱い小規模プラットフォームを利用することで，コミュニティ意識を醸成しようとします。特に暴力を助長するようなプライベートチャンネルやグループの可視性が低いことを利用し，代替プラットフォームを利用します。大規模化やウイルス化する前に，フォロワーを積極的に獲得しようとします。	外部からの質問やフィードバックに対し，明確で正確な情報と共感を持って対応できるよう，スタッフが研修プログラムを開発してください。外部のオーディエンスに対応するための十分なリソースを確保してください。 ソーシャルメディアチャネルにおけるコミュニティのガイドラインと行動規範を策定し，フォロワーに伝えてください。 迅速な情報共有と拡散を可能にするため，様々なコミュニケーションチャネルで活動するパートナーと連携してください。	様々なコミュニケーションチャネルを通じて，フォロワーや関係者からの質問，フィードバック，対話を促進してください。 ソーシャルメディアチャンネルにおけるコミュニティガイドラインと行動規範を策定し，スタッフとフォロワーに周知徹底してください。 管轄区域で適用される法律を周知徹底し，違法な手段を用いてこれらのチャンネルにアクセスした場合の結果を広く周知してください。	プラットフォームチェックを実施してください。 1）アカウント作成日などの詳細情報を確認 2）プロフィール写真と自己紹介 3）フォローしている他のサイトやアカウントを調査 4）投稿アクティビティを調査 5）ボットや荒らしの疑いのあるアカウントによる投稿がないか確認	管轄区域で適用される法律を周知徹底し，違法な手段を用いてこれらのチャンネルにアクセスした場合の結果を広く周知してください。

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.

種類	意図	防止 すべて	保護 政府関連	検出 すべて	対応
無防備な行為者への操作	二次拡散者が偽情報拡散者自身の発言を繰り返していることや，その発言が操作を意図していることに気づいていない場合が多く，著名人や組織を騙したり操作したりすることで，偽情報拡散を助長することを目的としています。感情に訴えるコンテンツを使用します。	経営陣に対し，個人および職場でのソーシャルメディア上の活動が，偽情報の拡散の標的となる可能性があることを教育してください。 フォロワーに対し，ソーシャルメディアでコンテンツを購読または共有する前に，情報源を確認し，評価するよう促してください。	啓発キャンペーンを実施し，積極的に偽情報に反論，または「プレバンキング」を行うことで，潜在的なオーディエンスを草の根の偽情報キャンペーンから保護してください。	プラットフォームチェックを実施してください。 1）アカウント作成日などの詳細を確認する 2）プロフィール写真と自己紹介 3）フォローしている他のサイトやアカウントを調べる 4）投稿アクティビティを調べる 5）コンテンツがボットや荒らしの疑いのあるアカウントによって投稿されていないか確認する。	職員に対し，個人および職場でのソーシャルメディア上の活動が偽情報の拡散の標的となる可能性があることを周知してください。 他のプラットフォームを利用して，政治問題や政策課題を明確にするメッセージを発信してください。

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.

C2PAガイダンスをサポートするツール

C2PA（コンテンツの出所と真正性に関する連合）規格をサポートするツールやライブラリはいくつかあり，特にコンテンツ真正性イニシアチブ（CAI）を通じて提供されています。

ツール	機能	使用時期
C2PAツール（コマンドラインユーティリティ）	C2PAマニフェストとメディアアセットを操作するための強力なCLIツール	マニフェストデータの読み取りと表示 マニフェストの添付と署名 サイドカーファイルの作成 信頼チェーンの検証 認証済みコンテンツを扱う開発者やメディア専門家に最適
CAIオープンソースSDK	C2PAをアプリケーションに統合するためのライブラリとツールのスイート:	JavaScript SDKを使用して，Webベースのコンテンツ認証情報の検証と表示を行ってください。 他のSDKで使用されるコア実装には，Rustライブラリを使用してください。 バックエンドまたはデスクトップアプリケーションには，プレリリース版のPython，Node.js，C++/Cライブラリを使用してください。 これらを使用することで，コンテンツ認証情報の作成，検証，表示が可能になります。
Web統合ツール	C2PAメタデータをWebサイトに埋め込み，表示するためのツール	データの出所をユーザーに表示する必要がある場合に便利です。デジタルアーティスト，ニュースルーム，プラットフォームにとって特に役立ちます。

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

マルチメディアコンテンツ認証ガイダンスのための最新および新興技術

ディープフェイク，誤情報，デジタル偽造が蔓延する時代において，マルチメディアコンテンツの認証技術はますます重要になっています。これらは，画像，動画，音声などのデジタルメディアの完全性，出所，真正性を検証するプロセスをカバーする技術です。様々な用途に活用できます。

デジタルアートや非代替トークン（NFT）の分野で所有権と独創性を証明するために使用できます。また，ユーザーが投稿した写真や動画の真正性を検証することが不可欠なジャーナリズム，ソーシャルメディアで改ざんされたコンテンツや偽造コンテンツを検出するためにも使用できます。さらに，法廷で使用されるデジタルメディアが改ざんされていないことを確認するという重要性が高まっている分野にも活用できます。

手法	機能	主要な規格とイニシアチブ
デジタル透かし	隠し情報（著作権やタイムスタンプなど）をメディアに直接埋め込みます。 脆弱（改ざんを検出）または堅牢（圧縮やサイズ変更に耐える）のいずれにも使用できます。	**JPEG Trust（ISO/IEC 19566 シリーズ）** JPEG 委員会（ISO/IEC JTC 1/SC 29/WG 1）によって開発されました。 デジタル画像の信頼性と来歴に焦点を当てています。 真正性を検証するための電子透かしとメタデータのサポートが含まれています。 **C2PA（Coalition for Content Provenance and Authenticity）** Adobe，Microsoft，BBC，Intel などの合同イニシアチブです。 デジタルコンテンツ出所メタデータと電子透かしのための標準化されたフレームワークを定義します。 ISO 規格ではありませんが，既に広く採用され，影響力を持っています。 **ISO/IEC 15444 (JPEG 2000)** 画像圧縮における電子透かしのオプションサポートが含まれています。 医療画像やデジタルシネマなど，高い忠実度とセキュリティが求められるアプリケーションで使用されます。 **AI透かしに関するITUとISOの連携** 国際電気通信連合（ITU）とISOは，以下の規格の開発に協力しています。 • AI生成コンテンツへの透かし • マルチメディアの真正性 • ディープフェイク検出

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)**

手法	機能	主要な規格とイニシアチブ
デジタル署名	暗号化技術を用いてメディアファイルに署名します。 改ざんがあった場合，署名は無効となり，整合性と真正性を保証します。	**公開鍵基盤（PKI）** • 秘密鍵を用いてコンテンツに署名し，公開鍵を用いて検証する，広く知られた信頼性の高い技術です。コンテンツが改ざんされていないことを確認し，署名者の身元を確認します。RSA，ECDSA，EdDSAといった一般的なアルゴリズムを使用します。 **分離署名と埋め込み署名** • 分離署名：署名はメディアファイル（例：.sigファイル）とは別に保存されます。 • 埋め込み署名：署名はメディアファイル（例：EXIFまたはXMPメタデータ）に埋め込まれます。 **ハッシュ＆署名** • メディアの暗号ハッシュが生成され，署名されます。 • 特に大容量ファイルの場合，効率的かつ安全です。 **タイムスタンプ** • コンテンツがいつ署名されたかを証明するために，署名に信頼できるタイムスタンプを追加します。 • 法的およびアーカイブ目的に役立ちます。 **ISO/IEC 9796 および 14888** デジタル署名スキームとメッセージ回復に関する規格 ハッシュとメタデータと組み合わせることで，マルチメディアに適用できます。 • PKI で公開鍵と ID を関連付けるために使用されます。 • セキュアメール，HTTPS，デジタルコンテンツ署名でよく使用されます。 **W3C 検証可能認証情報** コンテンツまたは ID に関するデジタル署名されたクレームのフレームワーク メディア作成者または発行者の真正性を検証するために使用できます。 **CAdES, XAdES, PAdES** **C2PA（コンテンツ出所）**

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.

手法	機能	主要な規格とイニシアチブ
ハッシュ化	ファイルに固有のハッシュ値を生成します。ファイルが変更されるとハッシュも変化します。 — 改ざん検出に役立ちます。	**暗号ハッシュ関数** ファイルの一意のフィンガープリントを生成するために使用される，標準的で安全なハッシュアルゴリズムです。SHA-256 / SHA-3（セキュアハッシュアルゴリズム）デジタル署名やブロックチェーンで広く使用されています。 メディアファイルに変更があると，ハッシュは全く異なるものになります。 NIST（FIPS 180-4）によって標準化されています。 知覚ハッシュ（pHash，aHash，dHash） コンテンツベースの識別に使用されます。サイズ変更や圧縮などの小さな変更にも耐性があります。 • pHash（知覚ハッシュ） • 画像または動画フレームのエッセンスを捉えます。 • 見た目が似ているメディアは，ハッシュも類似します。 • ほぼ重複したファイルや軽微な編集の検出に役立ちます。 • aHash（平均ハッシュ） • 平均ピクセル値に基づく簡略化された手法です。 • 高速ですが，pHash よりも堅牢性に欠けます。 • dHash（差分ハッシュ） • ピクセルの差分に基づきます。 • 構造変化の検出に適しています。 注記： MD5 / SHA-1 は，レガシーシステムや重要でない整合性チェックで依然として使用されています。しかし，安全性が低く，既知の脆弱性があるため，可能な限り使用を避けるべきです。 ビデオハッシュ手法： • フレームベースハッシュ: キーフレームに知覚的ハッシュを適用します。 • モーションハッシュ: 動きベクトルとシーンの変化を捉えます。 • 時間的ハッシュ: フレームの順序とタイミングを考慮します。 規格とフレームワーク： ISO/IEC 15938（MPEG-7）: マルチメディア内容記述インターフェース — 画像およびビデオ署名の記述子を含みます。 ISO/IEC 23000-19（MPEG-21 メディアバリューチェーン本体論）: コンテンツの識別と認証をサポートします。 【JSA注】 ISO/IEC 21000-19の誤り C2PA：暗号化ハッシュを用いてメタデータとコンテンツを安全に結び付けます。

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.

手法	機能	主要な規格とイニシアチブ
ブロックチェーンベースの認証	メディアのメタデータまたはハッシュをブロックチェーンに保存することで，改ざん防止と分散型の真正性記録を提供します。	**ブロックチェーン上のコンテンツハッシュ SHA-256** 権利とアクセスのためのスマートコントラクト Interplanetary File System や ＋Ethereum などのツールを使用することで，メディアをオフチェーンで保存し，ハッシュをオンチェーンで保持することで，ストレージコストを削減しながら整合性を維持できます。 **W3C 検証可能認証情報** コンテンツまたはアイデンティティに関するデジタル署名されたクレームの規格。多くの場合，分散検証のためにブロックチェーンと統合されています。 メディア制作者または発行者の真正性を検証するために使用できます。 **C2PA（コンテンツ出所と真正性のための連合）** ブロックチェーンネイティブではありませんが，ブロックチェーンと統合することで，変更不可能な出所追跡を実現できます。 **ISO/TC 307 – ブロックチェーン及び分散台帳技術** マルチメディア認証に関連するアイデンティティ，スマートコントラクト，データ整合性などの分野をカバーする，グローバルなブロックチェーン規格を策定している専門委員会
AIベースのフォレンジック	機械学習を用いて，ディープフェイクやスプライシングなどの改ざんの兆候を検出します。 照明，影，圧縮アーティファクトなどの不一致を分析できます。	**ディープフェイク検出** 畳み込みニューラルネットワーク（CNN）とトランスフォーマーを用いて，顔の動き，まばたき，照明，音声と映像の同期における不一致を探し，合成メディアを検出します。 アンチフォレンジック攻撃に対抗するため，GAN耐性を持つモデルがいくつか開発されています。 **スプライシングと改ざん検出** 圧縮アーティファクト，照明，または影の不整合を検出するために使用される手法。この手法では，マルチスケールCNNとアテンションメカニズムを用いて改ざんされた領域を特定します。 **OpenMFC（NIST）** NIST主導によるマルチメディアフォレンジックの課題とベンチマークの標準化イニシアチブ。 ディープフェイク検出，出所，アンチフォレンジックに焦点を当てています。 **C2PA（コンテンツの出所と真正性に関する連合）** 出所メタデータの埋め込みと検証のためのフレームワーク メタデータと出所分析 AIモデルは，メタデータと視覚コンテンツを相互参照することで異常を検出します。 多くの場合，トレーサビリティのためにブロックチェーンまたはC2PAフレームワークと統合されています。（ブロックチェーンのセクションを参照） **説明可能なAI（XAI）** AIの判断をフォレンジックアナリストや調査員が解釈できるようにすることで，信頼性を高めます。透明性が重要な場面で役立ちます。

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.

手法	機能	主要な規格とイニシアチブ
メタデータ分析	埋め込まれたメタデータ（EXIF，タイムスタンプ，GPS）を検査します。 不整合や編集の痕跡を明らかにできます。	**EXIF（Exchangeable Image File Format）** デジタル写真やフォレンジックで広く使用されている，画像ファイルにメタデータ（カメラモデル，日時，GPSなど）を保存するための規格です。 XMP（Extensible Metadata Platform）カスタムスキーマをサポートし，コンテンツ認証情報で使用されます。 Adobeによって開発され，様々なファイル形式にメタデータを埋め込むことができます。 **C2PA（Coalition for Content Provenance and Authenticity）** コンテンツの出所，編集履歴，所有権を追跡し，暗号署名されたメタデータをメディアに埋め込みます。 **MPEG-7（ISO/IEC 15938）** マルチメディアコンテンツ記述インターフェース。低レベルの機能（色，テクスチャ）と高レベルのセマンティクス（イベント，オブジェクト）の記述子を定義します。 **SWGDEベストプラクティス** デジタル証拠に関する科学作業グループは，デジタルビデオ認証におけるメタデータ分析のガイドラインを提供しています。 **ダブリンコアとIPTC** ジャーナリズムとデジタルライブラリで使用される標準化されたメタデータタグ付け。

Hash Comparison Chart

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

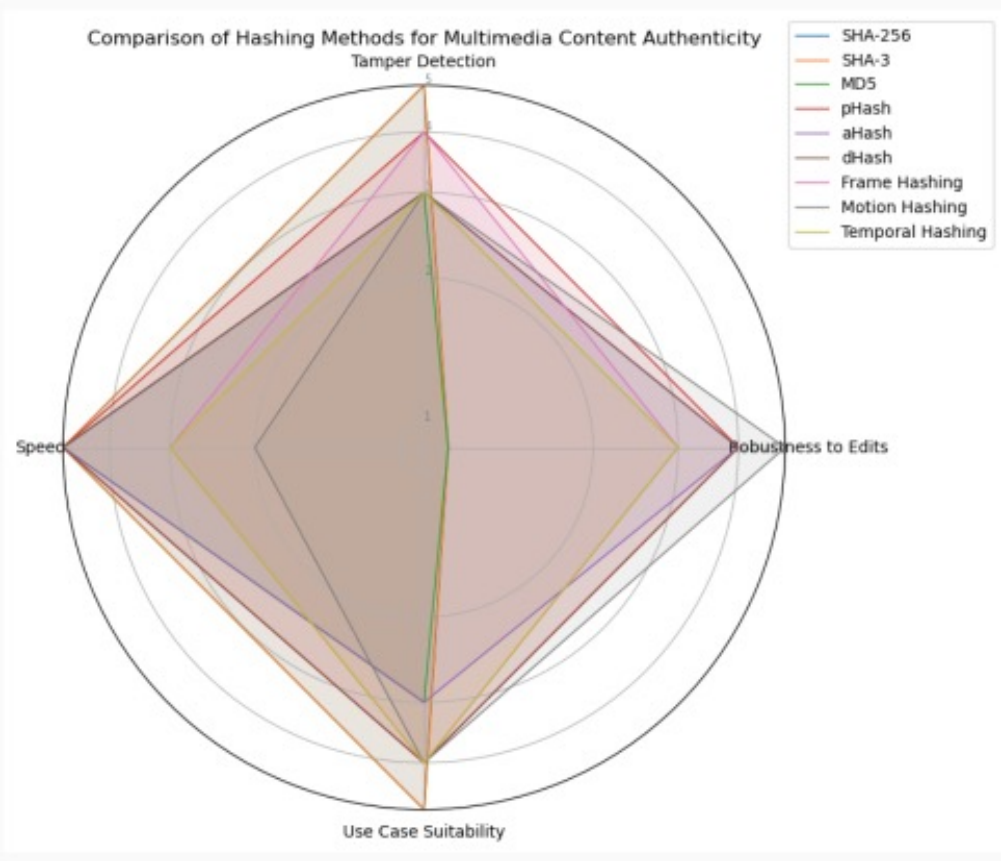
- Robustness to Edits
- Speed
- Tamper Detection
- Use Case Suitability

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



ハッシュ比較チャート

以下のチャートは，優先度に応じて異なるハッシュ手法を区別することを目的としています。

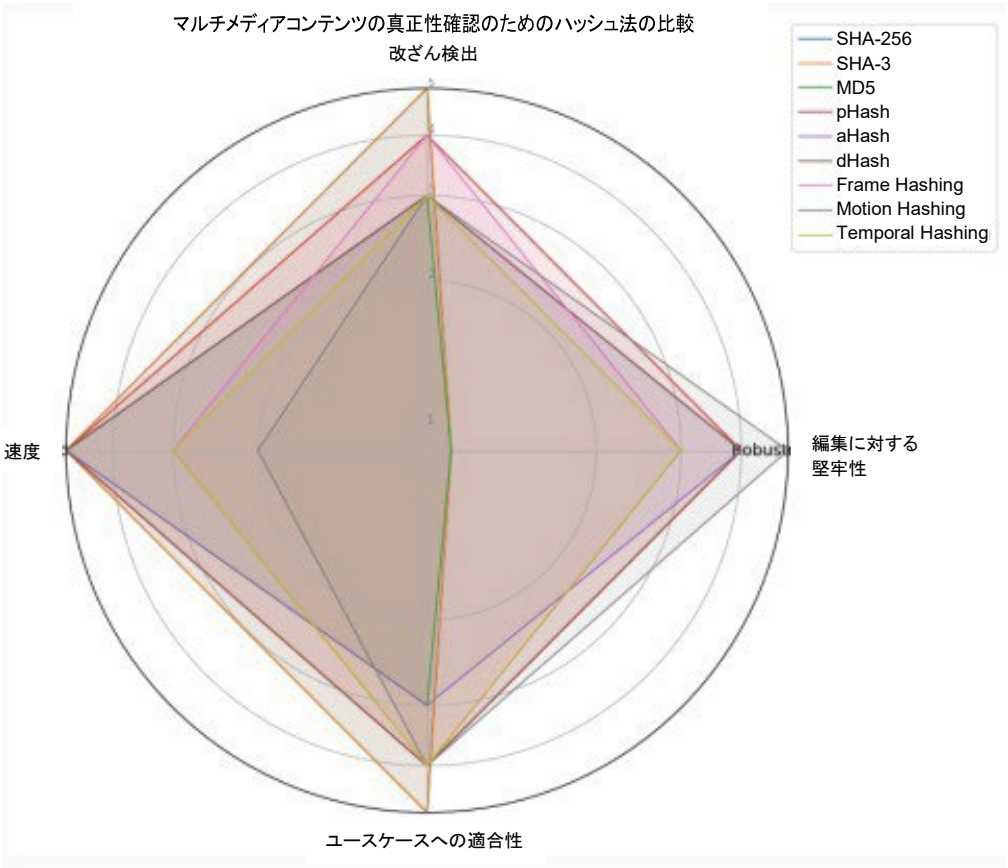
- 編集に対する堅牢性
- 速度
- 改ざん検出
- ユースケースへの適合性

解釈:

各線は異なるハッシュ手法を表しており，軸上で外側に向かうほど，そのカテゴリーにおけるパフォーマンスが優れていることに注意してください。

対象となるユースケースにおいて最も重要な基準は何かを判断してください。例えば，知覚的類似性の検出が主な関心事である場合，pHashの方が優れていることは明らかです。一方，速度と改ざん検出が優先される場合は，SHA-256が適しています。

マルチメディアコンテンツの真正性確認のためのハッシュ法の比較



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.	

一般チェックリスト

マルチメディアコンテンツの真正性確認チェックリストは，組織がデジタルコンテンツの完全性と出所を確認するのに役立ちます。

これには，コンテンツの出所，履歴，および変更の検証が含まれます。メタデータ，ソース情報，編集または改ざんの証拠に関するチェックは必ず実施してください。

トピック	チェック項目	チェック内容	結果
出典と出所			
1)	元の出典を確認する。	コンテンツの場所，時間，作成者を特定できるか。	
2)	メタデータを確認する。	カメラ設定，位置情報，タイムスタンプなど，埋め込まれている情報を確認する。	
3)	Reファイルの詳細を確認する。	ファイル名，バージョン，その他の属性を調べて，コンテンツの履歴に関する手がかりを探す。	
編集と操作			
1)	潜在的な改ざんを特定する:	レタッチ，トリミング，デジタル加工などの編集の痕跡を探す。	
2)	編集の影響を評価する:	変更がコンテンツの意味とコンテキストにどのような影響を与えるかを検討する。	
3)	編集履歴を記録する:	コンテンツに加えられた変更と，変更を行った人物を記録する。	
検証と妥当性確認			
1)	真正性ツールを使用する。	デジタルコンテンツの出典と履歴を検証できるソフトウェアまたはサービスを利用する。	
2)	専門家の指導は必要か？	必要に応じて，コンテンツの真正性またはメディアフォレンジックを専門とする専門家の指導を求める。	
3)	どのような規格を使用できるか？	業界規格として，コンテンツ真正性イニシアチブ（CAI）とC2PAを使用する。	

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.	

トピック	チェック項目	チェック内容	結果
透明性と開示			
1)	背景を提供する。	コンテンツがオリジナルか改変されたかを明確に示し，変更内容を説明する方法を確保する。	
2)	出典を正しく示す。	コンテンツの作成者，およびコンテンツの作成・編集に関わった個人または団体にクレジットが付与されていることを確認してください。	
3)	情報をオープンに共有する。	コンテンツの出所と履歴に関する関連情報を公開してください。	

透かしソリューション チェックリスト

透かしソリューションを選択する際に従うべきチェックリストを以下に示します。

このチェックリストの後に，著者によって検証されたソリューション名を示す表があります。その表は参考情報であり，推奨されるソリューションを主張するものではありません。このチェックリストを使用する際は，ソリューションの信頼性が高く，ご自身の用途に適していることを確認してください。

透かしソリューション チェックリスト		
選択前の設問		**回答**
	ソリューションは，画像，動画，音声など，さまざまな種類のコンテンツに対応していますか？	
	デモは閲覧可能ですか？	
	どの程度のレベルの保護を提供していますか？これは，コンテンツの具体的なニーズによって異なります。 今は必要なくても，脅威の増大に合わせて，より高度なセキュリティを提供するフォレンジック透かし機能を備えているかどうかを検討してください。	
操作性と使いやすさ	専門的な技術的知識を必要とせずに，コンテンツ制作者のワークフローにソリューションを簡単に統合できますか。	
	ワークフロープロセスに大きな変更を加えることなく，ソリューションを簡単にコンテンツ制作者のワークフローに統合できますか。	

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.	

透かしソリューション チェックリスト		
選定前の設問		**回答**
選定前の設問		回答
コスト	コストには何が含まれていますか？ 有料の「プラグイン」や「アドオン」は必要ですか？ アップデートは無料ですか？ 隠れたコストはありますか？	
ライセンス	ライセンスの詳細は？	
絞り込み後，以下のチェック項目に進みます：		
専門的仕様	リバースエンジニアリングや偽造に対するソリューションの堅牢性を確認するためのセキュリティレビューを実施します。	
	コンテンツ変更時の検出率を評価します。	
	コンテンツの種類全体にわたって，検知されにくいかどうかを分析および検証します。	
比較テスト	特定のコンテンツで並列テストを実施します。	
	業界規格の指標を使用してパフォーマンスを評価・測定します。	
拡張性，統合性，相互運用性	既存システムとの統合の容易さを評価します。	
	Perソリューションが現在のコンテンツ量に対応し，将来的に増加するコンテンツ量にも対応できるよう拡張できる能力を評価します。	

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.	

利用可能なソリューション

偏りや好みを示唆しないように，ソリューションはアルファベット順に表示されています。

ソリューション	用途	
DataPatrol	デバイスマーキングとWebマーキングに特化した多様なソリューションを提供します。	
Digimarc	様々なユースケースに対応するソリューションを提供し，GS1を使用しています。	
Digital Guardian/Fortra	幅広い透かしソリューションを提供します。	
Dropsend	機密文書向けの動的な透かしを提供します。	
Friend MTS	加入者ID透かしを含む，ライブスポーツやその他のエンターテイメント業界向けの透かしソリューションを提供します。	
Google DeepMind	AI生成コンテンツに透かしを入れるシステムであるSynthIDを提供します。	
IMATAG	フォレンジック透かしや監視サービスを含む，メディアおよび出版業界向けの透かしソリューションを提供します。	
MediaValet	メディア資産を保護するための透かしソリューションを提供します。画像の透かし入りレンディションの生成も含まれます。	
NAGRA	デジタルメディアとコンテンツを保護するためのフォレンジック透かしソリューションを提供します。	
NoisyPeak	オーディオおよびビジュアルコンテンツを保護するためのエンドツーエンドの透かしソリューションを提供します。既存のコンテンツに適用することも，追加のトランスコーディングやDRM保護を含めることもできます。フォレンジック保護とコンテンツ追跡。	
Synamedia	ContentArmorを含む，メディアおよびエンターテイメント向けのフォレンジック透かしソリューションを提供します。	

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チェックリスト		
領域	**設問**	**ガイダンス**
出典の検証：		
権威と信頼性：		
	著者，出版社，またはスポンサーは特定され，検証可能ですか？あるいは，その人物の身元を確認し，連絡を取ることができますか？	
	このアカウントをよく知っていますか？	
	それらのコンテンツと報道は過去に信頼できるものでしたか？	
	それらの専門知識または資格は検証できますか？	
	この出典は他の信頼できる出典によって引用されていますか？	
	この出典を信頼できる情報源は何かありますか？	
	アカウントにはどのような経歴情報が記載されていますか？	
	アカウントの主なナラティブ／議論のポイントは何ですか？	
	経歴情報がコンテンツの種類と矛盾していませんか？	例えば，コンテンツは高年齢層を対象としているのに，使用されている言葉遣いから，より若い年齢層によって作成または操作されたと示唆される場合などです。これは，都会的な語彙の使用によってしばしば識別できます。
	アップロード者の居住地を特定できますか？（下記のアカウント履歴を参照）	居住地はしばしば政治的な動機を示す指標となり，主張されている居住地，経歴情報，アップロード者の居住地の確認の間に矛盾がある場合に検出できます。
	他の場所にリンクしていますか？	
アカウント履歴（該当する場合）：		
	アカウントのアクティブ度は？	
	アップロード者はアカウントでどの程度アクティブですか？	
	過去にどのような種類のコンテンツをアップロードしましたか？	
	アカウント履歴に矛盾や警告サインはありますか？	

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チェックリスト		
領域	**設問**	**ガイダンス**
情報源のソーシャルネットワーク接続:		
	友人やフォロワーは誰ですか？	
	フォローしているのは誰ですか？	
	誰と交流していますか？	
	既知の偽情報チャンネルや人物と繋がっていますか？	
	より多くの情報を得るために，他のソーシャルネットワークで同じ名前／ユーザー名に関連付けられたアカウントを探してください。	実名が見つかった場合は，人物検索ツールを使用して，その人物の住所，メールアドレス，電話番号を見つけることができます。 Pipl.com White Pages Spokeo WebMii TwitterまたはFacebookの認証済みアカウントが実際に認証されているかどうかを確認するには，青いチェックマークにマウスオーバーしてください。アカウントがTwitterまたはFacebookによって認証されている場合は，「認証済みアカウント」または「認証済みページ」というポップアップが表示されます。 LinkedInで，その人物の職歴を確認してください。
コンテンツの検査:		
正確性と一貫性:		
	情報は他の信頼できる情報源で検証できますか？	
	タイムチェックを実施してください。	Wolfram Alphaなどのツールを使って，当日の天気などの詳細を検索し，その日の天気情報と，事案が発生したとされる場所を確認できます。地元の天気予報から気象状況を確認するのも良いでしょう。 現在見ているものより前に，同じ事案に関するコンテンツがないか確認してください。タイムスタンプを提供するツールや，Google，Tin Eye，YouTubeなどの動画・画像検索ツールを使うのも良いでしょう。 画像や動画の常識的な確認を無視せず，棚の時計，その日に放送されなかった番組を映しているテレビ画面，調査対象のコンテンツによるとまだ発生していない日付が記載されている新聞など，日時を裏付けるものや反証となるものを探してください。

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チェックリスト		
領域	**設問**	**ガイダンス**
	場所を確認してください。	Wolfram Alphaなどのツールを使えば，その日の天気などの詳細を検索し，その日の天気情報と，事案が発生したとされる場所を確認できます。画像や動画に表示されている気象状況が，Wolfram Alphaなどのツールで報告されたものと一致していることを確認してください。
		コンテンツに自動位置情報が含まれているかどうかを確認する必要があります。
		衛星画像や位置情報付き写真と比較できる参照点（道路標識や建物の看板など）を確認してください。車のナンバープレートが，提示された国以外の国で登録されているという異常な状況も探してください。広告看板は，その場所の言語で表示されていますか？ 主張されている位置情報を裏付ける，または反証できる特徴的な風景を探してください。スポーツスタジアムや大聖堂などを探すのも良いでしょう。 GoogleマップやGoogleストリートビューなど，無料で利用できるツールが数多くあります。
	調査には，主張を裏付ける十分な証拠が含まれていますか？	
	コンテンツ自体に矛盾や矛盾点はありますか？	
操作の兆候:		
画像		
	画像またはビデオは，加工または操作されているように見えますか？	ツールを使用して出所を確認してください。
	画像またはビデオは，付随するテキストと一致していますか？	ツールを使用して出所を確認してください。
	明らかな改変や歪曲はありますか？	リップシンクなどを確認してください。
ビデオ		
音声		

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チェックリスト		
領域	**設問**	**ガイダンス**
検索エンジンによる確認：		
	画像検索を実行し，画像が他のコンテキストで使用されているかどうかを確認します。	どのようなコンテキストで使用されていますか？扇動的，偏見的など，不適切な表現は含まれていませんか？
	検索エンジンを使用して，主張や情報の正確性を確認してください。	
ファクトチェック：		
	ファクトチェックウェブサイトで，コンテンツが既に検証されているかどうかを確認してください。	
	要に応じて，ファクトチェッカーにコンテンツを提出して検証を受けてください。	
コンテキストと裏付け：		
相互参照：		
	コンテンツが他の信頼できる情報源によって裏付けられているかどうかを確認してください。	上記の「正確性と一貫性」セクションを参照してください。
	偏りを避けるため，複数の情報源と照合して情報を検証してください。	上記の「正確性と一貫性」セクションを参照し，他の信頼できる情報源も確認してください。
適時性と関連性：		
	コンテンツは最新の出来事やトレンドに関連していますか？	タイムスタンプを確認してください。
	コンテンツは最新かつ正確ですか？	タイムスタンプを確認してください。
著者の視点と動機を分析する：		
	コンテンツの背後に偏見や意図はありますか？	使用されているナラティブや比喩は，特定の派閥やグループに共通しているかどうかを検討してください。
	このコンテンツを共有することで生じる可能性のある影響は何ですか？	共有の影響についてリスク評価を実施してください。

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

MMCAマトリックス

	必須	推奨

コンテンツの出所	強力	中程度	基礎的
	規格番号	規格番号	規格番号
	ISO 22144 コンテンツ認証情報	ISO 22144 コンテンツ認証情報	ISO 22144 コンテンツ認証情報
	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 情報技術－人工知能(AI) － AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能(AI) － AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能(AI) － AIシステムとAI支援の意思決定におけるバイアス
	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム
	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス
	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理
	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ
	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク
	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス

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

情報の信頼性と真正性	強力	中程度	基礎的
	規格番号	規格番号	規格番号
	ITU-TSG13 – ISO/IEC JTC 1/ SG 29 H.MMAUTH: マルチメディアコンテンツの認証のためのフレームワーク	ITU-TSG13 – ISO/IEC JTC 1/ SG 29 H.MMAUTH: マルチメディアコンテンツの認証のためのフレームワーク	ITU-TSG13 – ISO/IEC JTC 1/ SG 29 H.MMAUTH: マルチメディアコンテンツの認証のためのフレームワーク
	ISO/IEC TR 24028: 2020 情報技術－人工知能－人工知能における信頼性の概要	ISO/IEC TR 24028: 2020 情報技術－人工知能－人工知能における信頼性の概要	ISO/IEC TR 24028: 2020 情報技術－人工知能－人工知能における信頼性の概要
	ITU-T Y.3054 信頼に基づくメディアサービスのためのフレームワーク	ITU-T Y.3054 信頼に基づくメディアサービスのためのフレームワーク	ITU-T Y.3054 信頼に基づくメディアサービスのためのフレームワーク
	ISO/CD 22144 情報の真正性－コンテンツの信頼性	ISO/CD 22144 情報の真正性－コンテンツの信頼性	ISO/CD 22144 情報の真正性－コンテンツの信頼性
	ISO 24027 : 2021 情報技術－人工知能(AI)－ AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能(AI)－ AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能(AI)－ AIシステムとAI支援の意思決定におけるバイアス
	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム
	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス
	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理
	情報源真正性チェックリスト（ISAC）	情報源真正性チェックリスト（ISAC）	情報源真正性チェックリスト（ISAC）
	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ
	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク
	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス

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

透かし	強力	中程度	基礎的
	規格番号	規格番号	規格番号
	ISO/IEC 23078-1:2024 情報技術－デジタル出版物のためのDRM（デジタル著作権管理）技術の仕様	ISO/IEC 23078-1:2024 情報技術－デジタル出版物のためのDRM（デジタル著作権管理）技術の仕様	ISO/IEC 23078-1:2024 情報技術－デジタル出版物のためのDRM（デジタル著作権管理）技術の仕様
	SMPTE ST 2112-10:2020 コンテンツ識別子のオープンバインディング（OBID）	SMPTE ST 2112-10:2020 コンテンツ識別子のオープンバインディング（OBID）	SMPTE ST 2112-10:2020 コンテンツ識別子のオープンバインディング（OBID）
	2413-PLEN X.ig-dw: デジタル透かしの実装ガイドライン	2413-PLEN X.ig-dw: デジタル透かしの実装ガイドライン	2413-PLEN X.ig-dw: デジタル透かしの実装ガイドライン
	ISO/IEC TR 21000-11:2004 情報技術－マルチメディアフレームワーク（MPEG-21）－第11部：継続性連合技術の評価ツール	ISO/IEC TR 21000-11:2004 情報技術－マルチメディアフレームワーク（MPEG-21）－第11部：継続性連合技術の評価ツール	ISO/IEC TR 21000-11:2004 情報技術－マルチメディアフレームワーク（MPEG-21）－第11部：継続性連合技術の評価ツール
	IEEE P3361 デジタルコンテンツにおける電子透かし実装の堅牢性評価方法に関するIEEE規格案	IEEE P3361 デジタルコンテンツにおける電子透かし実装の堅牢性評価方法に関するIEEE規格案	IEEE P3361 デジタルコンテンツにおける電子透かし実装の堅牢性評価方法に関するIEEE規格案
	TR 104 032 人工知能（SAI）のセキュリティ保護	TR 104 032 人工知能（SAI）のセキュリティ保護	TR 104 032 人工知能（SAI）のセキュリティ保護
	ISO 24027 : 2021 情報技術－人工知能（AI）－AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能（AI）－AIシステムとAI支援の意思決定におけるバイアス	ISO 24027 : 2021 情報技術－人工知能（AI）－AIシステムとAI支援の意思決定におけるバイアス
	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム	ISO 42001 : 2023 情報技術－人工知能－マネジメントシステム
	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス	ISO 23894:2024 情報技術－人工知能－リスク管理に関するガイダンス

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

透かし	強力 規格番号	中程度 規格番号	基礎的 規格番号
	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理	ISO 12791:2024 情報技術－人工知能－分類および回帰機械学習タスクにおける不要なバイアスの処理
	管轄区域における法令遵守	管轄区域における法令遵守	管轄区域における法令遵守
	管轄区域における著作権法	管轄区域における著作権法	管轄区域における著作権法
	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ	ISO 27001: 2022 情報セキュリティ
	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク	GDPR またはその他のプライバシー規制／フレームワーク
	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス	OWAS セキュアコーディングプラクティス

ISBN 978-2-8399-4721-3

本文書は経済産業省の委託事業の成果です。

ISBN 978-2-8399-4721-3