Report

US POLICY

Synthetic Content

Exploring the Risks, Technical Approaches, and Regulatory Responses

October 2024

Authored By: **Jameson Spivack**, Senior Policy Analyst

FUTURE OF PRIVACY FORUM

Executive Summary

Generative AI (GenAI) enables the widespread creation of AI-generated or "synthetic" content, which is increasingly indistinguishable from human-generated or "authentic" content. While synthetic content predates GenAI, current GenAI tools' speed, processing power, and accessibility have given organizations and individuals the ability to create content for marketing campaigns, develop new medications, translate media into different languages, help individuals with speech impairments communicate, and more. At the same time, synthetic content—including images, video, audio, and text—that is indistinguishable from authentic content raises substantial risks for individuals, communities, and society.

As synthetic content becomes more common and more realistic, organizations are developing strategies to address the risks this content raises, including malicious impersonation, political disinformation and misinformation, and synthetic child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII). Policymakers, industry, academics, and civil society have also begun developing technical, organizational, and legal strategies for combating synthetic content's harms. In particular, these approaches typically include:

- Watermarking
- Provenance tracking
- Metadata recording
- Synthetic content labeling and disclosure
- Synthetic content detection
- Hashing and filtering
- Legal restrictions on deepfakes and impersonation

These techniques may help people make informed decisions about the content with which they interact online, and in some cases may support organizations' privacy and security commitments. However, they can also create privacy risks if they are implemented without safeguards for personal data. While no single approach will sufficiently mitigate the risks associated with synthetic content, stakeholders should consider the strengths and limitations of each when developing a comprehensive strategy for addressing synthetic content.

This report provides an overview of some of the risks synthetic content raises, explores the various approaches policymakers in the U.S. are taking to address these risks, and highlights some of these approaches' limitations, focusing on potential tradeoffs with privacy and security. The appendix provides further detail about the current major legislative and regulatory frameworks being proposed regarding synthetic content in the U.S.

Table of Contents

I. Introduction

While synthetic content is not necessarily new, the rapid growth of publicly available generative artificial intelligence (GenAI) tools has made it easier for people to create AI-generated visual, audio, and text content. As AI-generated or altered media becomes more difficult to distinguish from authentic content, and increasingly prevalent in the information ecosystem, policymaker attention has turned to addressing some of the risks this kind of content poses. Such risks include the use of synthetic content for malicious impersonation, political disinformation and misinformation, and synthetic child sexual abuse material (CSAM) and non-consensual intimate imagery (NCII).

In recognition of these risks, policymakers, industry, academia, and civil society have begun developing technical, organizational, and legal frameworks intended to mitigate some of the harms associated with synthetic content. There is no clear, widely-accepted line distinguishing "synthetic" content from "authentic" or non-synthetic content. Though this report largely reflects the more binary framing prominent among policymakers, technologists, and scholars, it should be noted that content's "synthetic" or "authentic" nature is closer to a spectrum.

U.S. lawmakers at the state and federal levels are exploring legislation requiring AI developers and deployers to implement techniques for addressing synthetic content's harms, such as watermarking, synthetic content detection and labeling, and data provenance tracking.[1] Responsive to the White House's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,[2] federal agencies such as the National Institute of Standards and Technology (NIST) are creating reports and guidance on synthetic content and authentication techniques.[3] Regulators at the Federal Trade Commission (FTC), Federal Communications Commission (FCC), and Federal Elections Commission (FEC) are also seeking to use their authorities to address impersonations, AI-generated spam calls, and deepfakes in political advertisements, in their respective domains.[4]

---

[1] *See* Appendix.
[2] *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, The White House (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.
[3] *NIST's Responsibilities Under the October 30, 2023 Executive Order*, NIST (Jul. 26, 2024), https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence.
[4] For the FTC's Supplemental Notice of Proposed Rulemaking (SNPRM) on the trade regulation rule regarding impersonation, see *FTC Proposes New Protections to Combat AI Impersonation of Individuals,* FTC (Feb. 15, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals. For the FCC's Notice of Proposed Rulemaking (NPRM) on AI-generated robocalls and robotexts, see *FCC Proposes First AI-Generated Robocall & Robotext Rules,* FCC (Aug. 7, 2024), https://www.fcc.gov/document/fcc-proposes-first-ai-generated-robocall-robotext-rules.

While safeguards for preventing harmful uses of synthetic content can support an organization's privacy and security efforts, they may also inadvertently create privacy risks, as well as tensions with the organization's data protection commitments and other legal obligations. Some techniques, such as those involving transparency or authentication, may reveal personal data, or require data be maintained indefinitely, potentially creating tradeoffs with privacy principles like data minimization. Some forms of synthetic content detection and identity authentication may also require more collection and analysis of personal data, including of private conversations. At the same time, a number of other factors may limit the effectiveness of techniques for combating harmful synthetic content, which should be considered when developing a holistic strategy for addressing these harms.

II. Synthetic content, or AI-generated content, can create or exacerbate risks.

The growth and widespread availability of GenAI tools has led to the proliferation of synthetic content, accompanied by a number of heightened privacy and safety risks.[5] *Synthetic content*, also called AI-generated content, refers to content—including text, audio, video, or other media—that has been created or "significantly altered" by algorithms.[6] Synthetic content is not inherently harmful, and can in fact increase productivity, help create more engaging virtual experiences, and personalize and improve medical diagnoses.[7] However, it can also be used maliciously to exacerbate existing risks related to fraud, manipulation, harassment, and more.[8] This is particularly true for women and people from marginalized communities, who often face

---

For the FCC's NPRM on the use of AI in political ads, see *FCC Proposes Disclosure Rules for the Use of AI in Political Ads*, FCC (Jul. 25, 2024), https://www.fcc.gov/document/fcc-proposes-disclosure-rules-use-ai-political-ads. For the FEC's consideration of a petition to regulate deceptive AI in political ads, see *Comments sought on amending regulation to include deliberately deceptive Artificial Intelligence in campaign ads*, FEC (Aug. 16, 2023), https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads. Note: the FEC has voted not to pursue this rulemaking.

[5] *Supra* 2.

[6] The U.S. federal government, following the White House's AI Executive Order, refers to synthetic content as that which is generated or "significantly altered" by AI. This framing has made its way into legislation as well. There is no consensus on what treatment of content constitutes a "significant" alteration; as such, the scope of synthetic content remains debated. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*, NIST (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf. *See also Supra* 2.

[7] Brenda Leong and Sara R. Jordan, *The Spectrum of Artificial Intelligence*, Future of Privacy Forum (Jun. 2023), https://fpf.org/wp-content/uploads/2023/07/FPF-AIEcosystem-Report-Jun23-R4-Digital.pdf.

[8] *Supra* 6.

greater harm when targeted with malicious synthetic content.[9] The harms that synthetic content can cause are highly context-dependent, and any approach to mitigating them must therefore be grounded in a thorough understanding of well-defined risks. This analysis focuses on harms broadly in the realm of privacy and security, setting aside other areas of risk, such as those related to markets and economics, intellectual property, and model performance.[10]

A. Malicious impersonation

The availability of software for generating "deepfakes"—synthetic content that appropriates a person's visual and/or audio likeness using AI—has made it easier to engage in impersonation and identity theft online. In fact, research suggests that impersonation or "manipulation of likeness" is one of the most common harmful uses of GenAI.[11] When a malicious actor engages in impersonation, two parties can be harmed: the person being scammed or defrauded as a result of the impersonation, and the person whose identity is stolen or misappropriated.[12] Both parties can experience financial, reputational, and emotional injury.[13] GenAI's continuing technical improvements have already, in many cases, made deepfakes nearly indistinguishable from non-synthetic content, leading to more convincing synthetic content. Often, deepfakes involve degrading content, meant to embarrass, defame, or bully others.[14] These attacks are disproportionately conducted on women and girls, with significant deleterious effects on their civil and political participation.[15]

B. Disinformation and misinformation

---

[9] Amber Ezzell, *Re: REG 2023-02 Artificial Intelligence in Campaign Ads,* Future of Privacy Forum (Oct. 16, 2023), https://fpf.org/wp-content/uploads/2023/10/Future-of-Privacy-Forum-FEC-Comment-on-AI-in-Campaign-Ads-October-16-2023.pdf. *See also* Danielle Keats Citron, *The Fight for Privacy: Protecting Dignity, Identity, and Love in the Digital Age*, W.W. Norton (2022), pg. 39.

[10] For a thorough examination of some of these other risks, see NIST's report "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." This report touches on risks related to weapons, confabulation, environmental impacts, anthropomorphism or emotional entanglement, IP, and value chain issues. *NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,* NIST (July 2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

[11] Nahema Marchal et al., *Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data*, arXiv (Jun. 5, 2024), https://arxiv.org/pdf/2406.13843.

[12] Jameson Spivack, Beth Do, and Angela Guo, *Re: Proposed Amendments to Trade Regulation Rule on Impersonation of Government and Businesses ("Impersonation SNPRM")*, Future of Privacy Forum (Apr. 29, 2024), https://fpf.org/wp-content/uploads/2024/04/FPF_FTC_SNPRM_Impersonation_Comment.pdf.

[13] Danielle Keats Citron and Daniel J. Solove, *Privacy Harms*, Boston University Law Review, Vol. 102 (2022), https://www.bu.edu/bulawreview/files/2022/04/CITRON-SOLOVE.pdf.

[14] *Supra* 7.

[15] PlanUSA, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0031*, NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0031.

GenAI has the potential to drastically increase the volume of disinformation and misinformation online.[16] Synthetic content's highly polished and convincing nature makes it more likely that more people believe the material is non-synthetic and/or accurate. Synthetic disinformation and misinformation can be particularly insidious when published without any acknowledgement that the content is synthetic, and when widely shared before users or platforms can label or fact-check the information appropriately. People may then further share the information in public and private networks, which can lead to even greater amounts of inaccurate or misleading information in the media ecosystem. Additionally, because GenAI tools may be trained on unfiltered datasets using reinforcement learning, false data—whether disinformation or misinformation—may be included in an AI model, leading to inaccurate and harmful outputs.[17]

1. *Elections and politics*

GenAI tools may make it easier to produce propaganda at scale. This includes content that contains inaccurate or misleading information, which could negatively impact public political understanding and engagement, as well as undermine election integrity. Malicious actors could use GenAI tools to imitate messages or behaviors from political candidates, parties, and interest groups to mislead voters or solicit funds or personal information.[18] For example, GenAI was used to create an AI-generated robocall impersonating President Biden and discouraging New Hampshire voters from voting in the 2024 primary,[19] as well as a false video of Ukrainian President Volodymyr Zelensky telling his soldiers to surrender.[20] Recently, former President Trump reposted false AI-generated images of Taylor Swift endorsing him for president to his followers on social media.[21] Malicious actors are also already using AI to create deepfake

---

[16] "Disinformation" refers to false information that has been shared *with the intent* to mislead or deceive people, or generally cause harm, whereas "misinformation" involves inaccurate or misleading information that is not necessarily intended to deceive. *Authenticating AI-Generated Content*, Information Technology Industry Council (Jan. 2024), https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf.
[17] *Id.*
[18] Daniel I. Weiner and Lawrence Norden, *Regulating AI Deepfakes and Synthetic Media in the Political Arena*, Brennan Center for Justice (Dec. 2023), https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena.
[19] Mekela Panditharatne, *Preparing to Fight AI-Backed Voter Suppression*, Brennan Center for Justice (Apr. 2024), https://www.brennancenter.org/our-work/research-reports/preparing-fight-ai-backed-voter-suppression.
[20] Bobby Allyn, *Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn*, NPR (Mar. 2022), https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.
[21] Elizabeth Wagmeister and Kate Sullivan, *Trump posts fake AI images of Taylor Swift and Swifties, falsely suggesting he has the singer's support,* CNN (Aug. 28, 2024), https://www.cnn.com/2024/08/19/politics/donald-trump-taylor-swift-ai/index.html.

non-consensual intimate imagery (NCII; see "Synthetic NCII" below), particularly of candidates who are women or people of color, to discourage their political and civic participation.[22]

2. *Health*

Synthetic health disinformation and misinformation can weaken public understanding of health issues and sow distrust in medical professionals and sciences, undermining both individual health decisions and community health outcomes more broadly.[23] Synthetic content may motivate individuals to try unproven and potentially harmful "remedies," or discourage them from seeking science-backed treatments, due to the content's highly realistic and believable nature.[24] Additionally, politically-motivated actors could create deepfakes of public health officials to spread disinformation, or create more realistic phishing attacks targeting patients. Synthetic disinformation and misinformation related to public health can even contain harmful advice cited from non-existent sources, which can easily confuse people who are unable to determine the origins of the content.

C. *Synthetic NCII*

Malicious actors are using GenAI models to create and disseminate synthetic non-consensual intimate imagery (NCII), a type of deepfake that alters photos and videos of real individuals and falsely portrays them, often in sexual manners, without their consent and often without their knowledge. The increasing accessibility of synthetic NCII creation tools can lead to sextortion, abuse, stalking, harassment, and humiliation, and disproportionately impacts women and girls.[25] It can often be difficult to determine who generated the fake images and if it was done with the knowledge or consent of the person portrayed, as this information is not always evident from the content itself, creating challenges for detecting and combating NCII online.[26] Some platforms have explicitly sought to monetize synthetic NCII, charging users a fee to "undress" their intended victim.[27]

---

[22] Coralie Kraft, *Trolls Used Her Face to Make Fake Porn. There Was Nothing She Could Do,* The New York Times (Jul. 31, 2024), https://www.nytimes.com/2024/07/31/magazine/sabrina-javellana-florida-politics-ai-porn.html.
[23] Tina Reed, *Deepfakes could supercharge health care's misinformation problem*, Axios (Nov. 2023), https://www.axios.com/2023/11/14/ai-deepfake-health-misinformation-fake-pictures-videos.
[24] *No laughing matter: navigating the perils of AI and medical misinformation*, Union for International Cancer Control (March 2024), https://www.uicc.org/news/no-laughing-matter-navigating-perils-ai-and-medical-misinformation.
[25] *Supra* 12.
[26] *Reducing Risks Posed by Synthetic Content*, NIST (April 2024), https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.
[27] Santiago Lakatos, *A Revealing Picture*, Graphika (December 2023), https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf.

D. Synthetic CSAM

As with the proliferation of NCII, GenAI tools can facilitate the creation and sharing of synthetic child sexual abuse material (CSAM), leading to significant emotional, reputational, and physical harm for minors.[28] Synthetic CSAM can depict either real children, whose publicly available, non-CSAM images are altered, or entirely fabricated images. Some AI image generation models that output synthetic CSAM are trained using datasets containing authentic CSAM, furthering the spread of harmful content.[29] Even when synthetic CSAM doesn't involve real children, it can still cause harm. Entirely synthetic CSAM that doesn't resemble actual children may divert law enforcement attention and resources away from investigating genuine CSAM.[30] The existence of synthetic CSAM may also contribute to the demand for CSAM, and may encourage behavior that leads to the harm of real children.[31] Notably, while CSAM is *prima facie* illegal in a number of jurisdictions globally, many are passing or considering legislation to clarify that synthetic CSAM is also covered under these laws.

E. Financial synthetic content scams

AI-generated text, voices, and videos can enhance the ability to create, target, and implement financial scams, including those involving malicious impersonation. In fact, phishing emails generated by a GenAI model were found to be more compelling than those written by a human,[32] which could lead to more effective and harmful scams. Scammers can use synthetic audio tools to clone people's voices in order to manipulate concerned family members, friends, or colleagues into believing someone they know is in danger. For instance, a common scheme involves a scammer claiming that a relative has been kidnapped and asking for a ransom (often called a "grandparent scam"), or posing as a person's boss to convince employees to take certain financial actions, such as buying gift cards or phone cards.[33] Malicious actors have committed substantial financial fraud while employing synthetic content, using deepfake videos of

---

[28] *Supra* 10.
[29] David Thiel, *Identifying and Eliminating CSAM in Generative ML Training Data and Models*, Stanford Digital Repository (2023). Available at https://purl.stanford.edu/kh752sm9123.
[30] *Supra* 10.
[31] Department of Justice, *Child Sexual Abuse Material*, https://www.justice.gov/d9/2023-06/child_sexual_abuse_material_2.pdf.
[32] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani, *AI Model GPT-3 (dis) informs us better than humans*, Science Advances, Vol. 9, Iss. 26 (Jun. 2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10306283.
[33] Jon Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*, Carnegie Endowment for International Peace (Jul. 2020), https://carnegieendowment.org/research/2020/07/deepfakes-and-synthetic-media-in-the-financial-system-assessing-threat-scenarios. *See also* Francesca Visser, *What is a Deepfake - and How Are They Being Used by Scammers?*, The Bureau of Investigative Journalism (Jun. 2024), https://www.thebureauinvestigates.com/stories/2024-03-07/what-is-a-deepfake-and-what-are-the-different-types.

co-workers to exfiltrate millions of dollars from corporate accounts, and impersonating deepfake victims to evade bank security measures.[34]

F. Discrimination

Synthetic content can exacerbate discrimination by generating outputs that reflect existing biases in training data, having a disparate impact on vulnerable and marginalized communities. AI models are trained on data that contain real-world biases, and these biases can be reproduced in the algorithm's results. For example, AI models could create content that reflects gender or racial stereotypes, or that creates negative associations for those with certain sexual orientations.[35] In addition to discrimination as a distinct harm, GenAI's other potential harms—such as disinformation, scams, and the other aforementioned harms—may discriminately impact members of vulnerable communities such as the elderly, non-native English speakers, and immigrants.[36]

G. Loss of trust in media

Widespread adoption of GenAI may make it challenging to determine what content is authentic, eroding the public's trust in media. Polarization and distrust in media institutions have increased in recent years, and documented cases of harmful online disinformation and misinformation continue to grow.[37] With more synthetic content in the information ecosystem, people may become more cynical about the media they encounter, or be hesitant to share news-related content out of fear that it's fake, leading to less confidence and participation in media.[38]

III. Policymakers, scholars, and technologists are creating frameworks for technical and organizational approaches to mitigating some of the risks associated with synthetic content.

---

[34] Heather Chen and Kathleen Magramo, *Finance worker pays out $25 million after video call with deepfake 'chief financial officer'*, CNN (Feb. 4, 2024), https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html. *See also* Emily Flitter and Stacy Cowley, *Voice Deepfakes Are Coming for Your Bank Balance*, The New York Times (Aug. 30, 2023), https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.

[35] *Challenging systematic prejudices: an investigation into bias against women and girls in large language models*, International Research Centre on Artificial Intelligence, UNESCO (2024), https://unesdoc.unesco.org/ark:/48223/pf0000388971.

[36] Grant Fergusson et al., *Generating Harms: Generative AI's Impact & Paths Forward*, Electronic Information Privacy Center (May 2023), https://epic.org/wp-content/uploads/2023/05/EPIC-Generative-AI-White-Paper-May2023.pdf.

[37] Minos Bantourakis, *How can we build trustworthy media ecosystems in the age of AI and declining trust?*, World Economic Forum (Oct. 2023), https://www.weforum.org/agenda/2023/10/news-media-literacy-trust-ai.

[38] *Supra* 7.

In recognition of the risks synthetic content poses, policymakers, scholars, and technologists are developing technical, organizational, and legal approaches for preventing or mitigating potential harms.[39] Many of these approaches are concerned with *authenticating* content, which involves verifying the source, history, and/or modifications of a piece of content.[40] These strategies help people determine whether content has been created or altered using GenAI, and provide transparency into the process by which it was created. While authentication techniques have been developed specifically for synthetic content, many argue that the same techniques can—and should—be adopted for non-synthetic content as well, to help people make more informed decisions about what media to trust in general.[41] Beyond authentication, another common approach is to place limitations on particular uses of synthetic content.

The following approaches can be grouped by several different criteria. They can be *direct* or *human-readable*, meaning they're intended to provide information that people can perceive and understand explicitly; or they can be *indirect* or *machine-readable*, meaning they're intended to be detected and analyzed by a machine, not a human.[42] These approaches can also be either *proactive*, applied intentionally for others to be able to learn more about the content; or *derived*, able to be applied regardless of whether the content creator intended for the content's history to be readable.[43] Whether a technique is human-readable or machine-readable, or proactive or derived, impacts the value it brings to the public, and its susceptibility to tampering.[44] Importantly, these approaches are not always separate or mutually exclusive, and often techniques will be combined to improve their effectiveness. For example, watermarking can be implemented by itself, or as part of provenance tracking.

A. Watermarking

[39] In January 2024 alone, state lawmakers introduced 101 bills addressing deepfakes. *BSA Analysis: States Intensify Work on AI Legislation*, BSA | The Software Alliance (Feb. 14, 2024), https://www.bsa.org/news-events/news/bsa-analysis-states-intensify-work-on-ai-legislation.
[40] Shayne Longpre et al., *Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?,* arXiv (Aug. 30, 2024), https://arxiv.org/pdf/2404.12691. *See also Supra* 16.
[41] Brad Smith, *Protecting the public from abusive AI-generated content* , Microsoft (Jul. 30, 2024), https://blogs.microsoft.com/on-the-issues/2024/07/30/protecting-the-public-from-abusive-ai-generated-content.
[42] *Supra* 6.
[43] *Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure,* Partnership on AI (Dec. 19, 2023), https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure.
[44] For example, while human-readable disclosure methods may provide more useful information to a casual observer about a piece of content, they are also more likely to be noticed and removed or altered by a malicious actor. *See Supra* 7.

Watermarking refers to the process of embedding information into content for the purpose of verifying the authenticity of the output, determining the identity or characteristics of the content, or establishing provenance (see "Provenance tracking" below).[45] Watermarking has long been used to deter piracy or track the spread of pirated content, though it is now being proposed as a way to help people determine the source and history of content. Digital watermarks can be either *overt* (human-readable), or *covert* (machine-readable), and can indicate the origin or source of content, as well as whether it's AI-generated.[46] While covert watermarks are more secure and harder to remove, they require a "decoder" to uncover them and extract their encoded information (see "Synthetic content detection" below),[47] and this decoder must be interoperable with the machine that embedded the watermark in order to work.[48] Overt watermarks, on the other hand, are immediately apparent and accessible, though may disrupt a piece of content or limit its use, and may also stand out as an obvious target for tampering.[49] Additionally, watermarks can be either *generic*, in which they identify a class of files rather than any individual person or transaction, or *individualized*, in which each watermark is unique and could reveal information about a person or behavior.[50]

B. Provenance tracking

Provenance tracking refers to recording and tracking the origins and history of content or data—also known as "provenance"—in order to determine its authenticity or quality.[51] Provenance data can be used to track where training data comes from, qualities of this data, what system generated a particular piece of content, and if or how it was altered by AI.[52] Provenance can be tracked in a number of ways, including embedding the information in digital watermarks (see "Watermarking" above) or through metadata recording (see "Metadata recording" below).[53] Like many technical approaches, provenance tracking is intended to improve transparency to help people better evaluate the content they encounter online, as well as to dissuade malicious actors

---

[45] *AI Output Disclosures: Use, Provenance, Adverse Incidents*, National Telecommunications and Information Administration (Mar. 27, 2024), https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-output-disclosures.
[46] *Supra* 6.
[47] *Id*.
[48] *Supra* 7.
[49] *Supra* 6.
[50] Individualized watermarks present more of a privacy risk than generic watermarks. *Privacy Principles for Digital Watermarking,* Center for Democracy and & Technology (May 2008), https://cdt.org/wp-content/uploads/copyright/20080529watermarking.pdf.
[51] *Supra* 6.
[52] *Supra* 16.
[53] *Supra* 6.

from creating or spreading CSAM or NCII by making it easier to track the original source of the content.[54]

Provenance can be tracked proactively, beginning at the time of the content's creation, or retroactively, after the content has already been created. Digital watermarking is a common way to proactively track provenance data, as it's easy to embed at the moment of content creation, whereas *authentication*—verifying claims made about the origin of particular content—is typically derived after content creation.[55] The Coalition for Content Provenance and Authenticity (C2PA), for example, has created an open metadata standard for cryptographically verifying the source and history of content, which organizations and creators can attach to their content.[56] C2PA bundles a piece of content with its provenance information and a cryptographic digital signature, keeping a permanent, tamper-resistant record of every time the content is modified.[57] Provenance tracking relies on an interoperable standard, such as C2PA's, that works for multiple file formats, and that the public generally understands.[58]

C. Metadata recording

Metadata recording refers to the process of tracking *metadata*—information about data or content itself, rather than its substance—for the purpose of authenticating the origins and history of content. There are several types of metadata: descriptive (e.g., file type, author), administrative (content source, ownership), technical (file type, size), structural (relationships between data elements), and provenance (origins).[59] Metadata can be internal (embedded in content) or external, can apply to any type of media,[60] and can indicate whether a piece of content is synthetic without harming the quality of the content in a way that humans would generally perceive.[61]

Importantly, metadata can be *signed*—stored using secure encryption and validated after it's been generated; or *unsigned*—not encrypted or validated, meaning it could be altered.[62] It's also possible to create digital fingerprints, which are "hashes" or codes generated to act as a unique identifier, to which metadata can be associated to identify harmful content like CSAM and NCII

---

[54] *Id*.
[55] *Supra* 45.
[56] Coalition for Content Provenance and Authenticity, https://c2pa.org.
[57] *Supra* 16.
[58] C2PA, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0030*, NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0030.
[59] *Supra* 6.
[60] *Id*.
[61] *Supra* 7.
[62] *Supra* 43.

(see "Hashing and filtering" below).[63] When metadata is recorded, people may be able to trace the history of content and determine whether it is synthetic.

D. Synthetic content labeling and disclosure

Synthetic content labeling and disclosures refer to methods of informing individuals that a given piece of content is synthetic, or that the individual is actively interacting with a GenAI system.[64] Disclosures may also provide additional relevant information about AI's role in the content generation, such as the particular AI system used to generate the content or information about the model's performance, training data, and capabilities.[65] This could include letting users know they're interacting with an AI chatbot; attaching content labels indicating if and how content has been generated or altered by AI; or publishing datasheets, model cards, or system cards containing information about the AI model.[66]

Labeling and disclosure are generally intended to improve people's ability to distinguish synthetic from non-synthetic content, and make informed decisions about content with which they interact. Disclosures can be human-readable, relying on visual or audio elements such disclaimers, tags, and nutrition labels; or they can be machine-readable, relying on techniques like covert watermarking or embedded code.[67] Human-readable disclaimers can take many forms, depending on the medium in question, and can be appended to content by either the creator or a third party.[68] Labeling primarily serves the goal of communicating how content was produced, though it can also—when paired with other mechanisms like fact-checking—aim to mitigate some of the harmful impacts of synthetic content.[69]

E. Synthetic content detection

"Synthetic content detection" is an umbrella term for a collection of tools and methods used to determine whether a given piece of content is synthetic. In contrast to techniques that label or

---

[63] *Supra* 6.
[64] *Supra* 45.
[65] *Id.*
[66] *Id.*
[67] *Supra* 7.
[68] For example, a content creator can indicate content is synthetic using titles, labels, pre-roll or interstitial disclosures, or annotation. Third parties could, in certain circumstances, label content as potentially synthetic through "speculation" mechanisms such as fact-checking or crowdsourced community notes. Tommy Shane, Emily Saltz, and Claire Leibowicz, *From deepfakes to TikTok filters: How do you label AI content?* First Draft (May 12, 2021), https://firstdraftnews.org/long-form-article/from-deepfakes-to-tiktok-filters-how-do-you-label-ai-content.
[69] Chloe Wittenberg et al., *Labeling AI-Generated Content: Promises, Perils, and Future Directions,* MIT Schwarzman College of Computing (Nov. 28, 2023), https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf.

embed information in content, detection techniques are external and typically analyze provenance or other characteristics to make determinations about whether content is synthetic. They can fall under three categories: automated content-based detection, which is applied after content has been generated; provenance data detection, such as identifying and analyzing watermarks embedded in content; and human-assisted detection, which explicitly involves humans in the detection process.[70] The specific detection tool that is appropriate for any given content will vary based on whether the content is an image, video, text, audio, or another content type.[71] When made publicly available, detection tools are intended to give people the ability to distinguish between synthetic and non-synthetic content they encounter.

F. Hashing and filtering

A method for preventing the spread of content like CSAM and NCII, whether synthetic or non-synthetic, is to create *hashes* that act as identifiers of particular harmful content, and to filter this content out of a given dataset or platform. First, a particular piece of content is identified and given a unique "hash" code, sometimes referred to as a "fingerprint," which may change when the content is altered. Then, the content can be compared to databases of hashes to identify instances of this hash—and the associated content—that can then be flagged or removed. This tends to work better in instances in which the content is able to be checked against the presumptive match, particularly since there are several methods to evade the system, and even a slight alteration of content will render a hash ineffective.[72] Hashes and their associated content can be identified and filtered out at various levels including in the training data, in the AI model input data that users intentionally prompt, or in the AI model output itself.[73]

G. Legal prohibitions on deepfakes and impersonation

In addition to the aforementioned technical and organizational approaches to addressing the harms of synthetic content, an increasingly common legal approach is to prohibit the creation of certain types of synthetic content or certain uses of that content—particularly deepfakes and similar material—and providing mechanisms for those who have been affected to seek relief. Prohibitions vary by legislation, but can include use in elections or political messaging (sometimes limited to within a certain time frame before and after elections), creation of CSAM or NCII, or more general "deceptive" uses. This approach doesn't relate to any particular tool or

---

[70] *Supra* 6.
[71] *Id.*
[72] Amie Stepanovich and Felicity Slater, *A Conversation on Privacy, Safety, and Security in Australia: Themes and Takeaways*, Future of Privacy Forum (Dec. 2023), https://fpf.org/wp-content/uploads/2023/12/A-Conversation-on-Privacy-Safety-and-Security-in-Australia-Themes-and-Takeaways.pdf.
[73] It's possible to create hashes of known, confirmed CSAM or NCII and track its spread across the Internet, deleting it whenever it appears. *Supra* 6, 43.

technique for detecting, identifying, or blocking synthetic content; rather, it bans certain conduct and provides for legal relief.[74] Legislative and regulatory prohibitions can apply to individuals who engage in prohibited behavior, or to platforms that distribute or fail to remove prohibited content.[75]

Approach	Description	Legislation examples
Watermarking	Embedding information into content for the purpose of verifying the authenticity of the output, determining the identity or characteristics of the content, or establishing provenance.	2024 legislation: California AB 3050,[76] California AB 3211,[77] U.S. SB 2765,[78] U.S. HB 7766[79]
Provenance tracking	Recording and tracking the "provenance" (origins and history) of content or data in order to determine its authenticity or quality.	2024 legislation: California AB 1791,[80] California SB 2885,[81] California AB 3050,[82] California AB 3211,[83] U.S. HB 7766[84]

[74] *See* Appendix.
[75] For example, in 2023 the FEC sought comment on a petition brought by the civil society organization Public Citizen, urging the Commission to clarify that its existing prohibition on the "fraudulent misrepresentation" of political candidates and parties included the use of "deliberately deceptive" AI in campaign ads. *Comments sought on amending regulation to include deliberately deceptive Artificial Intelligence in campaign ads*, FEC (Aug. 16, 2023), https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads. Note: the FEC eventually declined to pursue this rulemaking. *See* Ashley Gold, *Scoop: FEC won't act on AI in election ads this year*, Axios (Aug. 8, 2024), https://www.axios.com/pro/tech-policy/2024/08/08/fec-ai-election-advertising-no-action.
[76] *AB-3050 Artificial intelligence,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3050.
[77] *AB-3211 California Digital Content Provenance Standards*, California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3211.
[78] *S.2765 - Advisory for AI-Generated Content Act,* Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/2765/text?s=3&r=1&q=%7B%22search%22%3A%22s2765%22%7D.
[79] *H.R.7766 - Protecting Consumers from Deceptive AI Act,* Congress.gov, https://www.congress.gov/bill/118th-congress/house-bill/7766.
[80] *AB-1791 Digital content provenance,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB1791.
[81] *AB-2885 Artificial intelligence,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2885.
[82] *Supra* 77.
[83] *Supra* 78.
[84] *Supra* 80.

Metadata recording	The process of tracking metadata (higher-level information about data or content) for the purpose of authenticating the origins and history of content.	2024 legislation: California SB 942,[85] Massachusetts AB 4788,[86] U.S. HB 7766[87]
Synthetic content labeling and disclosure	Disclosing to individuals that a given piece of content is synthetic, or that a GenAI system is being used, and providing additional relevant information about AI's role in the content generation.	Enacted: Colorado SB 205 (AI use disclosure), Utah SB 149 (AI use disclosure) 2024 legislation: California SB 942,[88] California AB 2013,[89] California AB 3211,[90] Massachusetts AB 4788,[91] Pennsylvania SB 1044,[92] U.S. SB 2691,[93] U.S. HB 3831,[94] U.S. HB 7766,[95] U.S. AI Transparency in Elections Act[96]
Synthetic content detection	An umbrella term for a collection of tools and methods used to classify whether a given piece of content is synthetic.	2024 legislation: California SB 942[97]

[85] *SB-942 California AI Transparency Act*, California Legislative Information, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB942.
[86] *Bill HD.4788,* General Court of the Commonwealth of Massachusetts, https://malegislature.gov/Bills/193/HD4788.
[87] *Supra* 80.
[88] *Supra* 86.
[89] *AB-2013 Generative artificial intelligence: training data transparency,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.
[90] *Supra* 78.
[91] *Supra* 87.
[92] *S.B. 1044,* Pennsylvania General Assembly, https://www.legis.state.pa.us/cfdocs/billinfo/billinfo.cfm?syear=2023&sind=0&body=S&type=B&bn=1044.
[93] *S.2691 - AI Labeling Act of 2023,* Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/2691.
[94] *H.R.3831 - AI Disclosure Act of 2023,* Congress.gov, https://www.congress.gov/bill/118th-congress/house-bill/3831/text?s=1&r=1&q=%7B%22search%22%3A%22hr3831%22%7D.
[95] *Supra* 80.
[96] *S.3875 - AI Transparency in Elections Act of 2024*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/3875/text.
[97] *Supra* 86.

Hashing and filtering	Creating hashes, or identifiers, of particular harmful content and filtering this content out of a given dataset or platform.	2024 legislation: U.S. STOP CSAM Act of 2023[98]
Legal prohibitions on deepfakes and impersonation	Prohibiting certain uses of synthetic content—particularly deepfakes and similar content—and providing mechanisms for those who have been affected to seek relief.	Enacted: Tennessee ELVIS Act[99] 2024 legislation: U.S. DEFIANCE Act,[100] U.S. Protect Elections from Deceptive AI Act,[101] U.S. TAKE IT DOWN Act[102]

IV. Safeguards against synthetic content harms can both support and be in tension with privacy and security.

Technical, organizational, and legislative safeguards represent various approaches to mitigating the risks of synthetic content. These techniques may bolster privacy and security, not just by virtue of limiting the harmful effects of synthetic content like fraud and harassment, but also by improving mechanisms for compliance with privacy laws and user privacy preferences. At the same time, some techniques—particularly those involving transparency—may create tension with privacy and data protection principles, or face other limitations or downsides. Lawmakers and organizations seeking to implement effective safeguards while also protecting privacy should consider these tensions when developing strategies for addressing harmful synthetic content.

A. Techniques for addressing harmful synthetic content can support privacy and security.

In many cases, techniques meant to address the harmful impacts of synthetic content can better serve privacy and security objectives. Legal prohibitions on malicious impersonation and

[98] *S.1199 - STOP CSAM Act of 2023*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/1199/text. Note: this applies to CSAM broadly, and does not mention or distinguish synthetic CSAM.
[99] *HB 2091,* Tennessee General Assembly, https://wapp.capitol.tn.gov/apps/BillInfo/Default.aspx?BillNumber=HB2091.
[100] *S.3696 - DEFIANCE Act of 2024*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/3696.
[101] *S.2770 - Protect Elections from Deceptive AI Act*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/2770/text.
[102] *S.4569 - TAKE IT DOWN Act*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/4569.

deepfakes, most evidently, boost privacy by disincentivizing the creation and distribution of synthetic content that inherently violates individuals' privacy and personal autonomy, and by providing victims recourse and relief. Similarly, hashing and filtering known CSAM and NCII can, in particular circumstances, limit the spread of this inherently harmful and privacy-invasive synthetic content. Additionally, requiring synthetic content to be labeled or watermarked, or to provide provenance data, may help people identify fraudulent activity and avoid scams or security vulnerabilities. Data provenance tracking can also play an important role in ensuring compliance with users' privacy preferences, controlling access to and use of data, and protecting against data leakage.[103] For example, the Data & Trust Alliance's Data Provenance Standards, developed by a consortium of 19 organizations to improve data trustworthiness, provide that actors within the AI ecosystem should be able to indicate whether datasets contain personal data, and the level of sensitivity, which informs how the data may be used.[104]

B. *Techniques for combating harmful synthetic content can be in tension with privacy and security.*

Transparency techniques may be in tension with privacy and security if they contain or reveal personal data, or conflict with an organization's privacy or data protection principles or commitments.

1. *Transparency techniques can reveal personal data.*

Techniques intended to improve transparency, like metadata recording and provenance tracking, if implemented without safeguards, could potentially reveal sensitive personal data, or information about an individual's relationship to a piece of content.[105] For example, individualized watermarks could be used to monitor people's media habits or online behavior, which may include sensitive information, without user awareness.[106] Additionally, when paired with server-side logging of the prompts that individuals feed into a generative system, watermarks could reveal the content that identified individuals are generating.[107] Once personal data is

---

[103] Elisa Bertino et al., *A roadmap for privacy-enhanced secure data provenance,* Journal of Intelligent Information Systems, Vol. 43 (May 31, 2014), https://profsandhu.com/journals/misc/jiis_provenance_2014.pdf.
[104] *Data Provenance Standards*, Data & Trust Alliance, https://dataandtrustalliance.org/work/data-provenance-standards.
[105] *Supra* 6.
[106] Center for Democracy & Technology, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0029,* NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0029.
[107] Gustaf Björksten and Daniel Leufer, *Identifying Generative AI Content: When and How Watermarking Can Uphold Human Rights,* Access Now (September 2023), https://www.accessnow.org/wp-content/uploads/2023/09/Identifying-generative-AI-content-when-and-how-watermarking-can-help-uphold-human-rights.pdf.

collected and has been integrated into an AI model, it can be difficult to remove, as current methods for "unlearning" data aren't reliable.[108]

These techniques may create significant privacy risks for individuals while providing mechanisms for bad actors to easily thwart their benefits. Techniques for making authentication more resilient and robust—such as making these methods covert, or making it difficult to tamper with tracking data—can exacerbate the privacy risks.[109] On the other hand, widely available techniques for removing this data—such as stripping metadata from content—may render transparency mechanisms unhelpful or obsolete.[110]

The severity of relevant privacy risks may vary across use cases, and safeguards can be tailored to take these unique considerations into account. The risks posed by transparency techniques in the medical space differ from those arising in, for instance, social media applications or video games. In the medical context, the sensitive nature of health data necessitates additional caution when implementing transparency techniques for synthetic content. For example, watermarking could potentially unmask training data, and metadata recording could reveal patient identities if not properly de-identified.[111]

The mechanisms required to implement and enforce these techniques, meanwhile, may also involve privacy-invasive measures. For example, one proposed solution to the problem of AI-generated scam phone calls is the use of AI detection tools that alert people that they may be interacting with an AI agent impersonating a human.[112] Notably, the Federal Communications Commission (FCC) is exploring whether it should encourage the development of "real-time call detection, call alerting, and call blocking technologies" and/or develop rules regulating their use to combat robocalls.[113] Such programs would, in theory, analyze the content of a phone conversation in real-time, detect the presence of an AI-generated voice, and alert the called individual if AI is detected. This may protect the individual against potential AI-driven scams, but it also requires real-time collection and analysis of personal, and potentially sensitive, information.

---

[108] *Supra* 40.
[109] Public Knowledge, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0062,* NIST (Jun. 2, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0062.
[110] World Privacy Forum, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0063,* NIST (Jun. 2, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0063.
[111] AdvaMed Imaging Division, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0027,* NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0027.
[112] Paula Boyd and Jennifer Oberhausen, *Comments of Microsoft Corporation In the Matter of: Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts,* FCC, CG Docket No. 23-263 (Dec. 18, 2023), https://www.fcc.gov/ecfs/document/1219842001792/1.
[113] *Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts (NPRM)*, FCC, CG Docket No. 23-263 (Aug. 8, 2024), https://docs.fcc.gov/public/attachments/FCC-24-84A1.pdf.

Without safeguards, mass monitoring of private phone conversations by third parties—whether implemented solely by individuals or as part of mandated government regulation enforcement—raises significant privacy risks. Additionally, hashing and filtering can assist with stopping the spread of known CSAM, but may be less effective when messaging or other platforms encrypt content, a key privacy-enhancing technology.[114]

2. *Transparency techniques can conflict with other privacy and data protection principles.*

Methods for improving the transparency of synthetic content, while important, may also cause tension with existing privacy and data protection principles. Techniques like provenance tracking and metadata recording require collecting more data about a piece of content, potentially including personal data about how individuals interact with the content, which may conflict with data minimization mandates to collect as little data as possible. While watermarks, provenance data, or metadata on synthetic content don't always contain personal information, if they are designed to do so there may also be tension with individuals' statutory rights to control or delete their personal data. Similarly, transparency techniques may require data be kept longer than otherwise necessary for recordkeeping purposes, potentially clashing with data retention limitations and widely-accepted best practices. Given the potential tensions between synthetic content transparency techniques and privacy, some have called for technical experts to study these tradeoffs and develop privacy-preserving approaches and solutions.

C. *Other factors may limit the effectiveness of techniques for combating harmful synthetic content, or raise new problems.*

Ensuring that technical, organizational, and legal safeguards are as effective as possible requires recognition of any potential limitations or downsides and commitment to addressing those issues. First and foremost, no single approach is a panacea that can, by itself, tackle all relevant risks. Additionally, many of these techniques face logistical challenges that could, if not properly considered, impact their effectiveness or create new issues.

1. *Transparency techniques are not sufficient in isolation.*

The safeguards against harmful synthetic content promoted in current regulatory frameworks, while important, represent a relatively narrow, primarily technical set of potential approaches. Holistically addressing the potential harms posed by synthetic content requires a more comprehensive approach that should include technical, organizational, and legal safeguards.

[114] *Supra* 73.

First, the existing technical approaches are, by themselves, not yet effective enough to address the problems with which they're being tasked, and each comes with a number of tradeoffs.[115] For example, filtering and hashing content involves a tradeoff between filtering too little, resulting in more harmful content, or filtering too much, resulting in worse AI model performance.[116] Filtering out CSAM and NCII is also difficult to do if such content is included in training data, and the latter is particularly hard to detect given the near impossibility of judging consent in a piece of content, absent more information.[117]

Additionally, current authenticity techniques only authenticate the source or veracity of data, not other metadata like privacy or copyright information.[118] The techniques themselves vary in effectiveness depending on the type of content they're addressing (e.g., image, audio, text), as well as the medium on which they're present and the specific tools in question.[119] There may also be variance in how effective a given tool is across languages and cultures.[120] As such, the details of the particular technique being deployed, and the context in which it is used, heavily impacts its efficacy.

Mandating transparency tools without updating existing legal regimes regarding relevant risks like impersonation and fraud—which also exist outside of synthetic content—will only partially address the harms. In 2024, the FTC proposed new protections against impersonation, including AI-driven impersonation, that could allow the agency to combat scammers that appropriate other peoples' likeness to commit fraud, though the proposed updates also extend beyond fraud to implicate any impersonation of an individual, real or fictitious, that impacts commerce.[121] If more narrowly scoped, similar updates across the regulatory landscape and legal system—including

---

[115] Amy Cadagin, *Comments of the Messaging Malware Mobile Anti-Abuse Working Group (M3AAWG) on NIST AI 100-4, Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency,* M3AAWG (2024), https://www.m3aawg.org/sites/default/files/m3aawg_comments_on_nist_ai_100-4_reducing_risks_posed_by_synthetic_content_an_overview_of_technical_approaches_to_digital_content_transparencymay2024.pdf.

[116] Difficulties arise regardless of where in the content creation and distribution process filtering occurs. At the training data level, filtering out too little content may allow harmful content to slip through, whereas filtering too much will lead to poor model output. Filtering at the input data level relies on human data labels, and block lists, which require continuous updating and can be circumvented. Filtering at the output level is reactive, and hashing confirmed CSAM or NCII requires significant coordination between entities. Additionally, even detecting CSAM for the purposes of removal may be legally risky, as possessing the content is a crime. *Supra* 6, 107.

[117] *Supra* 6.

[118] *Supra* 40.

[119] *Supra* 6.

[120] *Supra* 107.

[121] *FTC Proposes New Protections to Combat AI Impersonation of Individuals,* FTC (Feb. 15, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals.

copyright, torts, and criminal law[122]—may help ensure there is relief for those impacted by harmful uses of synthetic content.[123]

2. *Transparency techniques can be easy to circumvent.*

In many cases, evading transparency tools is relatively easy for those with the knowledge and incentive to do so. Watermarks, for example, can be removed or altered, and are especially vulnerable when they are overt, as their visibility makes them an easy target for tampering.[124] Perhaps even more threateningly, watermarks and provenance data can also be forged, creating a false history for a piece of content's origins and lineage and placing a trust signal where trust is not warranted.[125] Not only may synthetic content appear as genuine, but non-synthetic content may also be flagged as synthetic, further eroding trust in the information landscape. Additionally, it may be easy to get around content filtering techniques such as keyword block lists, which tend to be simplistic and static. Block lists need to be continuously updated, and have trouble adapting to the evolving, ambiguous nature of language.[126] Similarly, if hashed content is altered, the hash is no longer effective, and motivated actors could evade filters.[127]

3. *Effective transparency techniques require standardization, interoperability, and coordination.*

A current lack of standardization, interoperability, and coordination creates challenges for the effectiveness of transparency techniques, and could adversely impact the integrity of the information environment. Research indicates that people are less trusting of content labeled as AI-generated, and the addition of labeling or watermarks to synthetic content increases user suspicion of the content.[128] However, without a standardized system for all content, synthetic and

---

[122] *Supra* 36.

[123] The Bipartisan Senate AI Working Group, convened by Sen. Chuck Schumer, also recommends that lawmakers consider legislation protecting against the unauthorized use of another person's likeness using AI. Chuck Schumer, Mike Rounds, Martin Heinrich, and Todd Young, *Driving U.S. Innovation in Artificial Intelligence*, The Bipartisan Senate AI Working Group (May 2024), https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

[124] *Supra* 7. *See also* Bob Gleichauf and Dan Geer, *Digital Watermarks Are Not Ready for Large Language Models,* Lawfare (Feb. 29, 2024), https://www.lawfaremedia.org/article/digital-watermarks-are-not-ready-for-large-language-models.

[125] Verance Corporation, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0034,* NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0034.

[126] *Supra* 6.

[127] *Supra* 73.

[128] Sacha Altay and Fabrizio Gilardi, *People Are Skeptical of Headlines Labeled as AI Generated, Even if True or Human Made, Because They Assume Full AI Automation,* PsyArXiv Preprints (Oct. 11, 2023), https://osf.io/preprints/psyarxiv/83k9r. *See also* Chiara Longoni et al., *News from Generative Artificial Intelligence is Believed Less,* FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (Jun. 20, 2022), https://dl.acm.org/doi/10.1145/3531146.3533077.

non-synthetic alike, good faith actors are more likely to label or watermark their synthetic content, and bad faith actors are not. The result is increased user suspicion of good faith content than of bad faith content.[129] If users can't assume that every piece of synthetic content is marked as such—or, conversely, that every piece of non-synthetic content is not—then making informed decisions about content is much more difficult.[130] Users also need to understand how these tools work in the first place, which requires greater investment in digital literacy and public education.[131] Transparency techniques geared towards synthetic content also don't address malicious or harmful uses of non-synthetic content, particularly given that people tend to scrutinize non-labeled content less than labeled content.[132]

Implementing these techniques effectively at scale also requires significant coordination between a range of actors and disparate, sometimes conflicting, systems. To effectively hash and filter synthetic CSAM and NCII, for example, social media platforms, AI developers, Internet service providers, law enforcement and advocates all need to work together.[133] Additionally, watermarking only works effectively if the watermark and the detector are aligned, which is not always the case,[134] and many of the existing transparency tools for AI models lack interoperability, threatening their overall efficacy and utility.[135] At the same time, if multiple watermarking systems become widely used—potentially including unreliable ones—there could be confusion about which systems are trustworthy.

Even when transparency techniques are interoperable, there is currently no universal standard for assigning responsibility for implementing and maintaining them, leading to uncertainty about the roles each actor in the ecosystem—such as developers, deployers, or users—should play.[136] There is also no consensus on when content that has been shaped by AI becomes "significantly" modified enough so as to require a label, which could create discrepancies in labeling and confusion among the public regarding what labels mean.[137] For instance, AI can be used to change the color or texture of a photo, as has been done with both AI-powered and non-AI-powered tools for decades. However, it's not clear what amount of modification to a piece

---

[129] *Supra* 70.
[130] *Supra* 16.
[131] Wiley, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0046,* NIST (Jun. 2, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0046.
[132] *Supra* 129.
[133] *Supra* 6.
[134] *Id*.
[135] *Supra* 40.
[136] *Supra* 7.
[137] The Alliance for Trust in AI, *Comment on FR Doc # 2024-09824, Comment ID NIST-2024-0001-0035,* NIST (May 31, 2024), https://www.regulations.gov/comment/NIST-2024-0001-0035.

of content should necessitate a label, or whether there's any substantive difference between equivalent levels of modification done by AI-powered tools compared to non-AI-powered tools.

D. Maintaining privacy and security for digital content transparency techniques.

A number of organizations have published recommendations for developing strategies to address harmful synthetic content. The following principles reflect some areas of common agreement between these organizations regarding creating transparency techniques that also preserve privacy and security.

Data minimization and purpose limitation

- Avoid including personally identifiable information (PII) in watermarks, data provenance, and content labels.[138]
- The amount of detail provided in provenance data should be tied to the use case.[139]
- Limit secondary uses of watermarking and other transparency techniques.[140]

Controls

- Include privacy considerations in data provenance tracking, in order to foster more responsible and compliant use of datasets.[141]
- Provide clear and conspicuous notice—of both the presence of synthetic content and the use of any transparency techniques.[142]
- Limit the ability for different entities to read, alter, and delete watermarks and provenance data.[143]
- Make processes available for accessing, correcting, and redacting PII.[144]

Collaboration

- Ensure techniques are interoperable and accessible.[145]
- Collaborate on research and best practices.[146]
- Improve public education and media literacy on synthetic content.[147]

[138] *Supra* 50, 16.
[139] *Supra* 16.
[140] *Supra* 50.
[141] *Supra* 105.
[142] *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action,* Partnership On AI (Feb. 27, 2023), https://syntheticmedia.partnershiponai.org. *Supra* 50.
[143] *Supra* 50.
[144] *C2PA Security Considerations,* C2PA Specifications, https://c2pa.org/specifications/specifications/1.0/security/Security_Considerations.html#_protection_of_personal_information. *Supra* 50.
[145] *Supra* 16, 40, 146.
[146] *Supra* 146.
[147] *Id.*

V. Conclusion

The increasing sophistication and spread of GenAI and synthetic content, while potentially beneficial across a variety of domains, may also contribute to harms related to disinformation and misinformation, non-consensual intimate imagery, fraud, and more. As a result, a range of stakeholders has begun developing technical, organizational, and legal approaches to mitigating some of the harms associated with synthetic content, often involving transparency and authentication. While these techniques, such as watermarking and provenance data tracking, can help support organizations' privacy objectives and foster safer online spaces, if they are implemented without safeguards they could also create privacy risks due to their potential use of personal data. Given these risks and other limitations, stakeholders implementing these techniques—including policymakers, industry, and civil society—should ensure their approach adequately addresses any potential shortcomings.

VI. Appendix: Regulatory Frameworks in the U.S.

The following information is accurate as of October 11, 2024.

A. Legislation: synthetic content transparency, authentication, and prohibitions

The following chart is a non-exhaustive list of some of the major legislation related to synthetic content transparency, authentication, and prohibition introduced in the U.S. in 2024.

Jurisdiction	Bill	Description	Status
Enacted			
California	SB942[148]	Providers of generative AI systems must make freely available to users a tool for detecting AI-generated content, and must disclose when content is synthetic. Does not apply to providers operating platforms with only non-user generated content.	Signed.
California	AB2013[149]	Generative AI system developers must publicly document information about the data used to train their system, and disclose whether synthetic data was or is used to develop the system.	Signed.
Tennessee	HB2091[150]	Ensuring Likeness Voice and Image Security Act (ELVIS) Act. Updates Tennessee's Protection of Personal Rights law to additionally prohibit the unauthorized use of a person's voice.	Signed.
Utah	SB149[151]	Any person using AI to interact with another person in connection with a regulated occupation (eg, requiring state licensure) must disclose when generative AI is being used in the provision of services. Excludes "synthetic data," along with all "de-identified data," from the definition of "personal data." Part of a more comprehensive AI bill.	Signed.

[148] *SB-942 California AI Transparency Act*, California Legislative Information, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB942.
[149] *AB-2013 Generative artificial intelligence: training data transparency,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.
[150] *HB 2091,* Tennessee General Assembly, https://wapp.capitol.tn.gov/apps/BillInfo/Default.aspx?BillNumber=HB2091.
[151] *S.B. 149 Artificial Intelligence Amendments,* Utah State Legislature, https://le.utah.gov/~2024/bills/static/SB0149.html.

Not enacted			
California	AB1791[152]	Social media platforms must redact personal provenance data from user-uploaded content (unless the user provides consent), but are prohibited from redacting system provenance data. Exception for personal provenance data that is copyright management information.	Did not pass.
California	AB3211[153]	Generative AI providers whose tools may be used to create highly realistic synthetic images, videos, and audio must embed provenance data into synthetic content produced using the tools. Generative AI providers must also make publicly available to users a provenance data detection tool. Recording device manufacturers must offer users the ability to embed provenance data in non-synthetic data they produce with the device. Large online social media platforms must use labels to disclose machine-readable provenance data detected in synthetic content distributed on its platform.	Did not pass.
Connecticut	SB2[154]	Any person using AI to interact directly with consumers must disclose to each consumer that they are interacting with AI, except if obvious. AI developers must label synthetic content in a machine-readable format. AI deployers must disclose when content is synthetic during first interaction with consumers. Part of a more comprehensive AI bill.	Did not pass.
Ohio	SB217[155]	AI systems must be programmed to embed non-removable watermarks in synthetic content. Prohibits the creation, reproduction, and publishing of synthetic simulated obscene material involving minors and impaired people, which must be removed from a platform within 24 hours of notification. Prohibits using a replica of someone's likeness to impersonate or defraud.	Referred to Senate Judiciary Committee. Legislative session ends 12/31/24.

[152] *AB-1791 Digital content provenance,* California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB1791.
[153] *AB-3211 California Digital Content Provenance Standards*, California Legislative Information, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3211.
[154] *Substitute for S.B. No. 2 - An Act Concerning Artificial Intelligence,* Connecticut General Assembly, https://www.cga.ct.gov/asp/cgabillstatus/cgabillstatus.asp?selBillType=Bill&bill_num=SB00002&which_year=2024.
[155] *Senate Bill 217,* Ohio Legislature, https://www.legislature.ohio.gov/legislation/135/SB217.

Pennsylvania	SB1044[156]	Publishing synthetic content without clear and conspicuous disclosure is an unfair and/or deceptive trade practice under existing law. Exemption for entities acting in good faith and without actual knowledge that content is synthetic.	Referred to Senate Communications and Technology Committee. Legislative session ends 11/30/24.
U.S.	S2765[157]	Advisory for AI-Generated Content Act. Makes it unlawful to create synthetic content without including a watermark, according to standards issued by the FTC, FCC, Attorney General, and Department of Homeland Security.	Referred to Senate Commerce, Science, and Transportation Committee.
U.S.	S3312[158]	Artificial Intelligence Research, Innovation, and Accountability Act of 2023 (AIRIA). Covered internet platforms are prohibited from operating generative AI systems unless the system informs each user, before interacting with synthetic content, about the use of AI in creating the content. Directs NIST to research and develop standards on content provenance and authenticity, as well as best practices for synthetic content detection. Part of a more comprehensive AI bill.	Reported favorably from Senate Commerce, Science, and Transportation Committee.
U.S.	S__[159]	Content Origin Protection and Integrity from Edited and Deepfaked Media (COPIED) Act. Directs NIST to develop standards for content provenance information, watermarking, and synthetic content detection, as well as cybersecurity measures. Generative AI tool providers must allow content owners to attach provenance information into content. Prohibits removing or tampering with content provenance information. Prohibits the unauthorized use of content with provenance to train AI	Introduced.

[156] *S.B. 1044,* Pennsylvania General Assembly, https://www.legis.state.pa.us/cfdocs/billinfo/billinfo.cfm?syear=2023&sind=0&body=S&type=B&bn=1044.
[157] *S.2765 - Advisory for AI-Generated Content Act,* Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/2765/text?s=3&r=1&q=%7B%22search%22%3A%22s2765%22%7D.
[158] *S.3312 - Artificial Intelligence Research, Innovation, and Accountability Act of 2023*, Congress.gov, https://www.congress.gov/bill/118th-congress/senate-bill/3312/text.
[159] *Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content,* U.S. Senate Committee on Commerce, Science & Transportation (Jul. 11, 2024), https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content.

		models or generate synthetic content.	
U.S.	HR7766[160]	Protecting Consumers from Deceptive AI Act. Directs NIST to create task forces to make recommendations on standards and guidelines for content provenance metadata, watermarking, and digital fingerprinting. Generative AI providers and covered online platforms muse ensure synthetic content includes disclosures that it is AI-generated. Directs the FTC to engage in rulemaking to enforce.	Referred to House Committee on Energy and Commerce Subcommittee on Innovation, Data, and Commerce.

B. Legislation: deepfakes and impersonation

The following is a non-exhaustive list of some of the major state-level legislation related to deepfakes and impersonation enacted or passed in the U.S. in 2024:

- Elections: Alabama HB172, Arizona HB2394, Arizona SB1359, California AB2355, California AB2655, California AB2839 (temporarily blocked),[161] Colorado HB1147, Delaware HB316, Florida HB919, Hawaii HB2687, Idaho HB664, Indiana HB1133, Louisiana HB154, Louisiana SB97 (vetoed),[162] Minnesota HB4772, Michigan SB2577, New Hampshire HB1596, New Mexico HB182, Oregon SB1571, Utah SB131, Wisconsin AB664
- CSAM: California AB1831, Florida SB1680, Idaho HB465, Florida SB1680, Idaho HB465, Indiana HB1047, Iowa SB2243, Kentucky HB207, Oklahoma HB3642, South Dakota SB79, Tennessee HB2163, Utah SB2163, Utah HB148, Utah HB238, Virginia SB731, Wisconsin SB314
- NCII: Alabama HB161, California SB926, California SB981, Idaho HB575, Louisiana SB6, Utah HB148, Utah SB66, Washington SB1999
- General: California AB1836 (prohibition of deceased person's synthetic impersonation), California AB2606 (digital replicas and contracts), Illinois HB4875 (consent for digital replicas), New Hampshire HB1432 (fraudulent use of deepfakes), New Hampshire HB1688 (state use of generative AI), Tennessee SB 2091 (property rights for personal likeness)

The following is a non-exhaustive list of some of the major federal legislation related to deepfakes and impersonation introduced in the U.S. in 2023-2024:

---

[160] *H.R.7766 - Protecting Consumers from Deceptive AI Act,* Congress.gov, https://www.congress.gov/bill/118th-congress/house-bill/7766.

[161] A federal judge issued a preliminary injunction temporarily blocking enforcement of AB2839, claiming the law violates the First Amendment. *See* Maxwell Zeff, *Judge blocks California's new AI law in case over Kamala Harris deepfake*, TechCrunch (Oct. 2, 2024), https://techcrunch.com/2024/10/02/judge-blocks-californias-new-ai-law-in-case-over-kamala-harris-deepfake-musk-reposted.

[162] Gov. Jeff Landry, *RE: Senate Bill Number 97 of the 2024 Regular Session by Senator Royce Duplessis* (Jun. 20, 2024), https://www.legis.la.gov/Legis/ViewDocument.aspx?d=1382564.

- Candidate Voice Fraud Prohibition Act (HR4611)
- Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability (DEEPFAKES Accountability) Act of 2023 (HR5586)
- DEFIANCE (Disrupt Explicit Forged Images and Non-Consensual Edits) Act of 2024 (S3696)
- Intimate Privacy Protection Act (HR9187)
- No Artificial Intelligence Fake Replicas And Unauthorized Duplications (No AI FRAUD) Act (HR6943)
- Nurture Originals, Foster Art, and Keep Entertainment Safe (NO FAKES) Act of 2024 (S4875)
- Protect Elections from Deceptive AI Act (S2770)
- Protect Victims of Digital Exploitation and Manipulation Act of 2024 (HR7567)
- Require the Exposure of AI–Led (REAL) Political Advertisements Act (S1596)
- Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks (TAKE IT DOWN) Act (S4569)
- Quashing Unwanted and Interruptive Electronic Telecommunications (QUIET) Act (HR7123)

C. *Regulation: federal agency action on synthetic content*

Agency	Action	Description
Federal Trade Commission (FTC)	Supplemental Notice of Proposed Rulemaking (SNPRM) to amend Trade Regulation Rule[163]	In March 2024, the FTC announced an SNPRM proposing to amend the Trade Regulation Rule on Impersonation of Government and Businesses to add a prohibition on the impersonation of individuals. The proposed rule would cover the use of AI to impersonate individuals.
Federal Communications Commission (FCC)	Notice of Proposed Rulemaking (NPRM)[164]	In July 2024, the FCC announced an NPRM on requiring broadcasters to provide on-air and written disclosure when AI-generated content is used in political ads on TV and radio.

[163] *Trade Regulation Rule on Impersonation of Government and Businesses: Supplemental notice of proposed rulemaking; request for public comment*, Federal Register (Mar. 1, 2024), https://www.federalregister.gov/documents/2024/03/01/2024-03793/trade-regulation-rule-on-impersonation-of-government-and-businesses.

[164] *Disclosure and Transparency of Artificial Intelligence-Generated Content in Political Advertisements: Proposed rule,* Federal Register (Aug. 5, 2024), https://www.federalregister.gov/documents/2024/08/05/2024-16977/disclosure-and-transparency-of-artificial-intelligence-generated-content-in-political-advertisements.

FCC	NPRM and Notice of Inquiry[165]	In August 2024, the FCC announced an NPRM on transparency requirements for AI-generated artificial or prerecorded voice messages, as well as AI-generated text messages ("robocalls" and "robotexts," respectively). The FCC also announced a Notice of Inquiry regarding the development of real-time AI detection tools for phone conversations, and potential privacy issues.
National Institute of Standards and Technology (NIST)	Draft publication,[166] memo,[167] and guidance[168]	In response to the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO), NIST developed a draft report on the risks posed by synthetic content, and technical approaches to digital content transparency. NIST also developed a memo being sent by the Department of Commerce, responsive to the EO, to the Director of the Office of Management and Budget and the Assistant to the President for National Security Affairs on standards and tools for synthetic content and provenance. NIST will develop guidance for synthetic content authentication by December 24, 2024.
Office of Management and Budget (OMB)	Guidance[169]	In response to the EO, OMB—in coordination with the Secretary of State, Secretary of Defense, Attorney General, Secretary of Commerce, Secretary of Homeland Security, and Director of National Intelligence—will issue guidance to federal agencies on labeling and authenticating content they produce or publish.
Federal Elections Commission (FEC)	Notification of Availability in response to	In August 2023, the FEC announced a Notification of Availability seeking public comment on a rulemaking petition filed by Public Citizen to amend existing

[165] *FCC Proposes First AI-Generated Robocall & Robotext Rules,* FCC, Docket No. 23-362 (Aug. 7, 2024), https://www.fcc.gov/document/fcc-proposes-first-ai-generated-robocall-robotext-rules-0.
[166] *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*, NIST (Apr. 2024), https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf.
[167] *NIST's Responsibilities Under the October 30, 2023 Executive Order*, NIST (Jul. 26, 2024), https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence.
[168] *Id*.
[169] *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, The White House (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

	rulemaking petition[170] (declined)[171]	regulations to clarify that deliberately deceptive AI in campaign ads is prohibited. In August 2024, the FEC announced a Notice of Disposition, declining to engage in rulemaking.

D. Bipartisan U.S. Senate AI Working Group Roadmap for Artificial Intelligence Policy

In May 2024, the Bipartisan Senate AI Working Group, led by Senate Majority Leader Chuck Schumer (D-NY) and composed of Senators Mike Rounds (R-SD), Martin Heinrich (D-NM), and Todd Young (R-IN), released "Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States" (AI Roadmap).[172] The AI Roadmap, published after a series of convenings with AI experts and relevant stakeholders, provides guidelines to lawmakers about specific AI issues that regulation should focus on. Below are recommendations the AI Roadmap makes regarding synthetic content, transparency, and authentication.

The AI Working Group encourages the relevant congressional committees to:

- Consider developing legislation to establish a coherent approach to public-facing transparency requirements for AI systems, while allowing use case specific requirements where necessary and beneficial, including best practices for when AI deployers should disclose that their products use AI, building on the ongoing federal effort in this space. If developed, the AI Working Group encourages the relevant committees to ensure these requirements align with any potential risk regime and do not inhibit innovation.
- Review forthcoming reports from the executive branch related to establishing provenance of digital content, for both synthetic and non-synthetic content.
- Consider developing legislation that incentivizes providers of software products using generative AI and hardware products such as cameras and microphones to provide content provenance information and to consider the need for legislation that requires or incentivizes online platforms to maintain access to that content provenance information.

---

[170] *Comments sought on amending regulation to include deliberately deceptive Artificial Intelligence in campaign ads*, FEC (Aug. 16, 2023), https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads.

[171] Sean J. Cooksey, Allen J. Dickerson, and James E. "Trey" Trainor, III, *RE: REG 2023-02 (Artificial Intelligence in Campaign Ads) – Draft NOD*, FEC (Aug. 8, 2024), https://www.fec.gov/resources/cms-content/documents/mtgdoc-24-29-A.pdf. *See also Artificial Intelligence in Campaign Ads: Notification of disposition of Petition for Rulemaking*, Federal Register (Sep. 26, 2024), https://www.federalregister.gov/documents/2024/09/26/2024-21979/artificial-intelligence-in-campaign-ads.

[172] Chuck Schumer, Mike Rounds, Martin Heinrich, and Todd Young, *Driving U.S. Innovation in Artificial Intelligence*, The Bipartisan Senate AI Working Group (May 2024), https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

The AI Working Group also encourages online platforms to voluntarily display content provenance information, when available, and to determine how to best display this provenance information by default to end users.

- Consider whether there is a need for legislation that protects against the unauthorized use of one's name, image, likeness, and voice, consistent with First Amendment principles, as it relates to AI. Legislation in this area should consider the impacts of novel synthetic content on professional content creators of digital media, victims of non-consensual distribution of intimate images, victims of fraud, and other individuals or entities that are negatively affected by the widespread availability of synthetic content.

The AI Working Group encourages the relevant committees and AI developers and deployers to:

- Advance effective watermarking and digital content provenance as it relates to AI-generated or AI-augmented election content.

The AI Working Group encourages AI deployers and content providers to:

- Implement robust protections in advance of the upcoming election to mitigate AI-generated content that is objectively false, while still protecting First Amendment rights.

Acknowledgements

The author would like to thank those who contributed to, reviewed, and provided feedback on this report, including Anne Flanagan, Tatiana Rice, Sonia Saini, Amie Stepanovich, and John Verdi.

If you have any questions, please contact us at info@fpf.org.

Disclaimer: This report is for informational purposes only and should not be used as legal advice.

1350 EYE STREET NW | SUITE 350 | WASHINGTON, DC 20005

info@fpf.org | FPF.ORG