MERCATUS SPECIAL STUDY

THE DEEPFAKE CHALLENGE
TARGETED AI POLICY SOLUTIONS FOR STATES

Dean W. Ball, *Mercatus Center*

MERCATUS.ORG

MERCATUS CENTER
George Mason University

Dean W. Ball, *The Deepfake Challenge: Targeted AI Policy Solutions for States* (Mercatus Special Study, Mercatus Center at George Mason University, October 2024).

Abstract

This study focuses on AI content authenticity issues such as deepfakes. It analyzes the evidence about the extent and nature of AI-enabled malicious content, finding it to be a serious, though often overstated, problem. It then analyzes how existing state and federal laws apply, asking centrally, "Is it already a crime to distribute malicious deepfakes?" The study finds that it often is a crime, but not always under some state common and statutory law. Thus, it suggests, a targeted legislative approach could be taken by either states or the federal government to deal with the problem.

JEL codes: L86, K24, O33, D86, Z18

Keywords: AI content manipulation, AI deepfakes, AI-generated content, AI-generated fraud, AI misinformation, AI political misinformation, AI safety and policy, content authenticity, deepfake criminal liability, deepfake detection tools, deepfake legal framework, deepfake regulation, deepfake state laws, deepfake technology challenges, digital content provenance, digital watermarking, generative AI law enforcement, generative AI policy, metadata authenticity, synthetic content legislation, synthetic media regulation, technological standards for AI, watermarking standards

The views expressed in Mercatus Special Studies are the authors' and do not represent official positions of the Mercatus Center or George Mason University.

Technologists, policymakers, journalists, and members of the general public have expressed a wide range of AI-related concerns—everything from algorithmic bias to the extinction of the human species. But the concern that has been top of mind for policymakers, judging from the number of legislative proposals, is deepfakes. State legislative bodies have introduced more than 300 deepfake-related bills in the 2024 session alone, and several dozen have been enacted.

These bills have taken many different approaches. Some are narrow, focused on nonconsensual sexual imagery or political communications. Others are far broader, imposing sweeping mandates on social media platforms, generative AI providers, and even camera manufacturers.

Just as the legislative proposals have varied, so too has the technical feasibility of many of these bills. This problem is particularly significant for bills that require watermarks to validate the authenticity or provenance of digital content shared online. Although numerous technical watermarking standards have been proposed, many have serious flaws, which this report will explore. Further technological breakthroughs are always possible; nonetheless, the problem of proving the authorial authenticity of content has been a challenge for all of written human history, and it is heretofore unsolved. Whether validating the authenticity of every piece of digital content shared on the internet will *ever* be possible is therefore questionable. At the very least, one should have measured expectations and assume incremental rather than total progress.

Deepfakes, like many other aspects of generative AI, present policymakers and society with new challenges. The solution to those challenges is not obvious a priori, and finding the optimal set of laws to grapple with deepfakes will be a discovery process. One of the key findings of this report is that many existing state laws have gaps that may make it difficult for victims of deepfake-enabled abuse to seek legal redress. In many states, then, a new law is required. This issue would be addressed best with a federal law, but given the uncertainty and slowness of the congressional process, state lawmakers will feel a legitimate need to act now. This report lays out a road map that could apply at either the state or federal level, but given the high likelihood that state legislatures will adopt laws faster than Congress, it is intended primarily for state government officials.

To be successful in crafting AI-related laws, state governments must avoid mandates that are impossible to achieve technologically. Such mandates create the *illusion* of safety but not the reality of it. Instead, policymakers must grapple with this issue with realistic expectations about what laws can plausibly accomplish. Namely, this report recommends focusing on post hoc enforcement against people who provably distribute deepfake content with malicious intent rather than on ex ante laws intended to stop deepfakes from being possible in the first place. No law is likely to "solve" the problem of deceptive and malicious AI-generated content, and trying to eliminate such content altogether is likely to create more problems than it solves. Yet the right laws can make meaningful progress, which is all that can be expected when grappling with novel sociotechnical problems.

This special study will examine the nature and extent of problems with synthetic content, explore the gaps in existing legal frameworks with respect to deepfakes, outline the current state of deepfake-related legislation, outline the technological solutions that have been proposed, and propose a framework for deepfake legislation that can counter deepfakes without overreaching.

Is AI-Generated Content a Problem?

Many AI risks, including risks that legislation has been drafted to mitigate, are speculative rather than demonstrated. Society has not observed AI models that can autonomously design bioweapons or execute devastating cyberattacks, as some in the AI safety community have warned may one day come to fruition. The AI models observed so far instead resemble many early technologies: helpful tools with more than their fair share of quirks and downsides. Many of the direst AI-related risks remain theoretical. This fact is not a reason to discount those risks, but it is a reason to be skeptical of imposing costly regulations that implicitly assume that those theories are true.[1]

Deepfakes, however, are not theoretical concerns. Almost everyone who has spent time on the internet since generative AI systems such as Midjourney,[2]

[1] Arvind Narayanan and Sayash Kapoor, "AI Existential Risk Probabilities Are Too Unreliable to Inform Policy," *AI Snake Oil*, July 26, 2024.

[2] Midjourney (website), https://www.midjourney.com/home.

Stable Diffusion,[3] and ChatGPT[4] first became mainstream two years ago has seen a deepfake. At the same time, the impact of deepfakes and other synthetic content has been less than some had expected, with some academic experts suggesting that deepfakes may "destroy democracy."[5] Two years on, researchers now have sufficient experience to make some early assessments about the actual, versus perceived, risks of synthetic content.

Synthetically generated content has undoubtedly been used for malicious purposes, including the following:

- Nonconsensual sexual material
- Fraud
- Misinformation and propaganda

Unfortunately, comprehensive data on this diffuse and fast-moving subject is difficult to find. Sumsub, a company involved in online fraud detection, reported in late 2023 that its internal statistics suggest a 1,000 percent increase in the number of detected deepfake images or videos globally and a 1,740 percent increase in North America.[6] On the whole, however, these deepfakes still account for a relatively small portion of overall incidents of online fraud, rising from 0.2 percent in 2022 to 2.6 percent in 2023.[7]

Anecdotally, several high-profile incidents have occurred since 2022 involving AI-generated content. In early 2024, a finance worker in Hong Kong was tricked into transferring $25 million to criminals using AI to impersonate his employer's chief financial officer on a video call.[8] A similar incident occurred in China, although the money was ultimately recovered.[9]

AI-generated material has also, rather predictably, found its way into political discourse around the world. During the 2024 Democratic Party primary in New Hampshire, a political consultant used an AI-generated clone of President

[3] Stability.AI, "Stable Diffusion 3" (website), https://stability.ai/news/stable-diffusion-3.

[4] OpenAI, "Introducing ChatGPT," November 30, 2022, https://openai.com/index/chatgpt/.

[5] Richard Painter, "Deepfake 2024: Will *Citizens United* and Artificial Intelligence Together Destroy Representative Democracy?," *Journal of National Security Law and Policy* 14 (2023): 121. See also Michael Waldman, "The Danger of Deepfakes to Democracy," *The Briefing*, Brennan Center for Justice, March 26, 2024.

[6] Sumsub, "AI-Generated Fraud and Deepfakes to Grow," *The Sumsuber,* December 27, 2023, https://sumsub.com/blog/sumsub-experts-top-kyc-trends-2024/#ai-generated-fraud-and-deepfakes-to-grow.

[7] Keepnet Labs, "Deepfake Statistics and Trends About Cyber Threats 2024," April 16, 2024, https://keepnetlabs.com/blog/deepfake-statistics-and-trends-about-cyber-threats-2024.

[8] Dylan Butts, "Deepfake Scams Have Robbed Companies of Millions. Experts Warn It Could Get Worse," CNBC, May 27 2024.

[9] Yang Zekan, "Deepfake Video Scams Prompt Police Warning," *China Daily,* March 6, 2024.

Joe Biden's voice to place calls to potential voters urging them not to vote.[10] In a tight presidential election in Slovakia last year, AI-generated audio of one of the leading candidates discussing plans to rig the election went viral on social media. The clip was immediately debunked by media outlets, but they were unable to disseminate this information because of a law in Slovakia prohibiting media coverage of politics in the 48 hours before an election.[11] The candidate in question ultimately lost, although whether the deepfake itself was the monocausal explanation for the loss is, of course, not clear.

The type of crime outlined here is not new. Deepfakes, synthetic misinformation, and associated malicious behavior have been prevalent online for many years. Although new tools to generate such content are surely more capable than earlier versions, they do not constitute a fundamentally new category of threat. Instead, generative AI represents a continuation of trends that long predate the rise of Midjourney, Stable Diffusion, ChatGPT, and similar products.[12]

Indeed, given the quality and widespread availability of these tools, it is, in fact, surprising that misuse is not more prevalent, particularly with regard to political misinformation. Many experts from a diverse range of fields warned with a high degree of confidence that 2024 would be a landmark year for AI-generated political misinformation, given the large number of elections being held across the world. Although such material has been propagated, whether it has had a meaningful impact on the information environment is not yet obvious.

In just a few weeks, the United States saw a wide range of fast-moving, high-profile political events, such as President Biden's withdrawal from the presidential election and the first assassination attempt on former President Trump. Yet, by and large, AI-generated misinformation did not predominate in social media feeds or other coverage of those events. Although the information climate related to those events was replete with misinformation, it was generally of a far more ancient variety: human beings lying, sharing incomplete or out-of-context information inadvertently, or jumping to conspiracy theories. These incidents are surely problems, but assuming that they are new problems or that they are susceptible to legislative solutions would be foolish.

[10] Shannon Bond, "A Political Consultant Faces Charges and Fines for Biden Deepfake Robocalls," NPR, May 23, 2024.

[11] Morgan Meaker, "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy," *Wired,* October 3, 2023.

[12] Tim Hwang, "Deepfakes: A Grounded Threat Assessment," Center for Security and Emerging Technology, July 2020.

Beyond overtly malicious uses of AI-generated content, other, largely inadvertent problems are associated with the rise of synthetic content. Perhaps chief among those problems is the spread of AI-generated articles of dubious quality and veracity.[13] Such content often is created by websites seeking to gain top placement in search engine results, a process known as search engine optimization (SEO). To some extent, it is a continuation of a cat-and-mouse game that search engine providers such as Google and Microsoft have been engaged in with so-called content farms for many years.[14] Regardless, such material can disrupt an internet user's access to high-quality information from the web.

In addition to this problem, however, the rise of synthetic content online also burdens—ironically enough—generative AI companies themselves. These companies collect enormous amounts of data from the internet—indeed, so much as to approximate the entire statistical distribution of text available online.

As AI-generated data come to represent a greater share of the training data for future generative AI models, the quality of those future models may be degraded. This phenomenon occurs because training on synthetic data—data created by other generative AI models—can lead to a phenomenon known as "model collapse," in which the model becomes unable to perform as intended.[15] Although synthetic data are used by all leading AI researchers and companies to *improve* future models, it is done in a highly deliberate and sophisticated manner.[16] Simply training off the outputs of random, often relatively low-performing generative models is likely to lead to bad results. Thus, generative AI companies have an incentive to identify and remove such content from their training datasets and, therefore, an incentive to create technical standards that help achieve that goal.

Other harms from deepfakes have yet to be demonstrated but are foreseeable enough—and directly implicate key government functions—that they merit attention from policymakers today. Perhaps chief among them is the use of AI-generated data in the judicial system. Both parties to a civil or criminal dispute have an obvious incentive to use AI-generated data to prove or disprove

[13] Robert Mariani, "The Dead Internet to Come," *The New Atlantis,* Summer 2023.

[14] "The Cat and Mouse Game of SEO," Bozzell, February 18, 2009, https://bozell.com/thinking/articles/the-cat-mouse-game-of-seo/.

[15] Ilia Shumailov et al., "AI Models Collapse When Trained on Recursively Generated Data," *Nature* 631 (July 2024): 755–59.

[16] Ruibo Liu et al., "Best Practices and Lessons Learned on Synthetic Data for Language Models," April 9, 2024. See also OpenAI: Dylan Royan Almeida, "Synthetic Data Generation (Part 1)," *OpenAI Cookbook,* April 9, 2024, https://cookbook.openai.com/examples/sdg1; Microsoft: Marah Abdin et al., "Phi-3 Technical Report," May 2024; Anthropic: Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," December 2022; and Nvidia: "Nemotron-4 340B Technical Report," June 2024.

the charges in question. Courts have only just begun to update their evidentiary standards to reflect this potential threat to a well-functioning legal system.[17] Given that it is one of the bedrocks of American society, addressing this threat to the legal system is an issue deserving of urgent attention.

In conclusion, then, synthetic content generated by AI models has resulted in material harms. Those harms have generally been less than was predicted by experts when the generative AI wave began in 2022—in some cases, far less. Still, deepfakes clearly present a foreseeable risk. Do existing laws address those risks effectively, or is a new law needed?

Gaps in Existing Law

Worth noting is that the suspected bad actors in many of the cases described in the previous section were charged with crimes under existing laws in their respective countries.[18] This point underscores an important fact about AI policy: many of the foreseeable harms are already crimes under current law, obviating the need for new criminal statutes or regulations.[19] That assertion is partially true in the case of deepfakes, but a close examination of common and statutory law in many states reveals gaps in many existing, pre-generative AI legal frameworks.

Many states have common law or statutory frameworks protecting the right of privacy—a right against undue intrusions into an individual's property or private affairs—and the right of publicity, which deals with a person's ability to control how their identity is used in public settings. Those frameworks vary in implementation by state and sometimes have substantial gaps, which are summarized in the following list. For a more detailed treatment of this subject, see the US Copyright Office's July 2024 report on digital replicas.[20]

- Some states do not provide for a right of privacy or publicity at all; others provide for only one or the other.
- Some states protect only certain classes, such as public figures (celebrities, politicians, journalists, etc.) rather than the entire population.

[17] Chief Justice John G. Roberts Jr., "2023 Year-End Report on the Federal Judiciary," December 31, 2023. See also "Artificial Intelligence and the Courts: Materials for Judges," American Association for the Advancement of Science, September 2022.

[18] See Zekan, "Deepfake Video Scams Prompt Police Warning," and Bond, "A Political Consultant Faces Charges."

[19] Howe Whitman III, Daniel Wiser Jr., and Dean Woodley Ball, "How to Worry, Not Panic, About Artificial Intelligence," *National Affairs* 60 (Summer 2024).

[20] United States Copyright Office, "Copyright and Artificial Intelligence Part 1: Digital Replicas," July 2024.

- Many states provide a right to sue for misuse of an individual's likeness only if the misuse took place in a commercial context.
- Some state laws do not protect against the unauthorized misuse of an individual's voice.
- Some preexisting state laws are too broad for use with AI, creating a right to sue even if no malicious intent exists or in the event that AI-generated content merely resembles a person rather than directly duplicating their likeness.

One can readily imagine scenarios involving AI-generated outputs in which each gap would create unintended problems. The first four gaps identified could prevent a wronged individual from seeking legal redress. The last gap could allow overzealous litigants to create a chilling effect on the use of AI for positive economic and social ends. Thus, legislation is clearly needed—at least in many states—to address those gaps. Another solution would be a federal law, but given the uncertainty of legislation passing at the federal level, states are poised to take action more quickly.

With that scenario in mind, this report turns to a survey of current state-based deepfake legislation.

The Landscape of State-Based Deepfake Legislation

Deepfake laws take one of two broad approaches:

- Ex ante regulation imposes requirements on generative AI developers, social media platforms, and related firms to prevent the dissemination of deceptive AI-generated content.
- Post hoc laws create civil or criminal liability for users who disseminate certain kinds of AI-generated content (nonconsensual sexually explicit material, deepfakes of politicians running for elected office, etc.).

Virtually all deepfake-related legislation passed by state governments to date has been post hoc laws.[21] Laws of this sort are easier to enforce because they operate by relieving a demonstrated harm rather than preventing that harm in the first place.

However, post hoc laws have their own complexities, especially as they pertain to political content. The First Amendment has broad protections for

[21] Ballotpedia Artificial Intelligence Deepfake Legislation Tracker (database), https://legislation.ballotpedia.org/ai-deepfakes/home.

political speech, and drawing the line between satirical political speech and deceptive content can be difficult in practice.[22] How courts will interpret deepfake laws as they apply to political speech remains to be seen, but given recent rulings from the Supreme Court and other federal courts, broad protections applied to the use of AI would not be surprising.[23]

More broadly, policymakers should have measured expectations about the legal viability of deepfake/content authenticity laws made to mitigate against "misinformation." Adjudicating what constitutes "misinformation" is often difficult, particularly when doing so is most important: in closely watched events that are unfolding in real time. That difficulty is part of the reason the First Amendment specifies that "Congress shall make *no* law" affecting freedom of speech and why the First Amendment has been incorporated to cover state laws as well.[24] Although important exceptions exist (false advertising, perjury, defamation, fraud, etc.), Americans in general have a right to make false statements in public, even to millions of people, under the First Amendment.[25] They will likely retain that right regardless of whether they use generative AI to do so.

In addition, many of the post hoc laws that states have passed do not meaningfully address the gaps in existing law just highlighted—or they do so only partially. For example, many states have passed laws that broadly protect politicians from deepfakes while doing little to protect other citizens.[26] Others have focused only on nonconsensual sexual material.[27] Robust protections are needed instead

[22] See, for example, Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988), https://supreme.justia.com/cases/federal/us/485/46/.

[23] Moody v. NetChoice, LLC, 603 U.S. ____, No. 22–277, 34 F. 4th 1196 (2024), https://www.supremecourt.gov/opinions/23pdf/22-277_d18f.pdf.

[24] Gitlow v. New York, 268 U.S. 652 (1925), https://supreme.justia.com/cases/federal/us/268/652/.

[25] United States v. Alvarez, 567 U.S. 709 (2012), https://supreme.justia.com/cases/federal/us/567/709/. See also McIntyre v. Ohio Elections Commission, 514 U.S. 334 (1995), https://supreme.justia.com/cases/federal/us/514/334/.

[26] See Alabama House Bill 172: "Crimes & Offenses, Provides Criminal & Civil Penalties for Distribution of Materially Deceptive Media Intended to Influence an Election," https://legiscan.com/AL/text/HB172/2024; Arizona Senate Bill 1359: "Election Communications; Deepfakes; Prohibition," https://legiscan.com/AZ/text/SB1359/2024; Hawaii Senate Bill 2687: "Relating to Elections," https://legiscan.com/HI/text/SB2687/2024; Indiana House Bill 1133: "Use of Digitally Altered Media in Elections," https://legiscan.com/IN/text/HB1133/id/2869748/Indiana-2024-HB1133-Introduced.pdf. The full list of enacted bills is available here: Ballotopedia (database), https://legislation.ballotpedia.org/ai-deepfakes/search?status=Enacted&category=Political%20communications&session=2024&page=1.

[27] See the following: Alabama House Bill 161: "Crimes & Offenses, Prohibits a Person from Creating a Private Image Without Consent," https://legiscan.com/AL/text/HB161/2024; Iowa House Bill 2240: "A Bill for an Act Relating to Harassment by the Dissemination, Publishing, Distribution, or Posting of a Visual Depiction Showing Another Person in a State of Full or Partial Nudity or Engaged in a Sex Act that Has Been Altered to Falsely Depict Another Person, and Making Penalties Applicable"

for *all* citizens whose likeness is digitally replicated by an actor with malicious intent.

Other post hoc legislation errs by attaching liability to the developer of the generative AI system rather than exclusively on the distributor of malicious deepfake content. For example, Tennessee's Ensuring Likeness, Voice, and Image Security (ELVIS) Act, passed in the spring of 2024, states the following:

> A person is liable to a civil action if the person distributes, transmits, or otherwise makes available an algorithm, software, tool, or other technology, service, or device, the primary purpose or function of which is the production of an individual's photograph, voice, or likeness without authorization.[28]

In practice, determining whether a generative AI model's "primary purpose" is the production of digital replicas of a specific person is challenging. Many such tools have built-in safety filters to prevent users from creating deepfakes of specific individuals. Yet in another sense, the primary purpose of almost all generative AI systems is, indeed, to create believably human-generated content. Rather than punishing toolmakers who release AI models, post hoc laws should focus on punishing users who create *and distribute* deepfake content for provably malicious purposes.

In general, post hoc laws have the benefit of being easier to enforce and, when crafted well, responding to demonstrated harms that must be proven in the judicial system.

Ex ante deepfake laws, on the other hand, are significantly more challenging to craft and execute effectively. They impose requirements designed to stop the dissemination of deceptive AI-generated content—a laudable goal whose feasibility is currently unclear. By requiring generative AI companies, websites and apps, and other firms to comply with those standards, states may create onerous regulatory burdens. They may also create other unintended consequences. Although no ex ante laws have been passed as of this writing, several have been proposed. A case study of one such bill will suffice to demonstrate the challenges.[29]

(formerly HF 2048), https://legiscan.com/IA/text/HF2240/id/2909187. The full list of enacted bills is available here: Ballotopedia (database), https://legislation.ballotpedia.org/ai-deepfakes/search?status=Enacted&category=Pornographic%20material&session=2024&page=1.

[28] Tennessee House Bill 2091: "AN ACT to Amend Tennessee Code Annotated, Title 39, Chapter 14, Part 1 and Title 47, Relative to the Protection of Personal Rights," https://legiscan.com/TN/text/HB2091/id/2900923.

[29] For more, see Dean W. Ball, "California's Other Big AI Bill," *Hyperdimensional,* July 29, 2024, https://www.hyperdimensional.co/p/californias-other-big-ai-bill.

Case study: California Provenance, Authenticity, and Watermarking Standards Act (AB 3211)

AB 3211 was a bill introduced in the 2024 legislative session by California Assemblymember Buffy Wicks.[30] Though the bill did not ultimately become law, it had significant momentum, unanimously passing in the Assembly and winning support from OpenAI.[31] As the most ambitious watermarking bill introduced in any state legislature, it is a case study in the difficulties of *ex ante* AI content regulation.

AB 3211 aimed to address the challenges associated with AI-generated content by mandating rigorous standards for watermarking and labeling of synthetic media. The bill proposed that all AI-generated content have "difficult to remove" watermarks, though "difficult to remove" was not defined. In addition, it demanded that websites with more than two million California users label synthetic and nonsynthetic content and maintain databases of potentially deceptive material. Furthermore, makers of recording devices were required to offer watermarking options for the content they capture.

The intent behind AB 3211 was to combat misinformation and the misuse of AI-generated content by making it easier to distinguish between human- and AI-produced media. However, the bill faced several significant implementation challenges. First, the technical feasibility of creating watermarks that are "difficult to remove" could be, depending on how that term is defined, a major hurdle. Existing standards have proved ineffective because they can be easily removed or altered, thus failing to prevent deception—a topic that will be addressed in detail later in this paper.

Moreover, AB 3211 imposed strict and broad requirements on AI developers, regardless of their size or the nature of their products. Requirements included the need for AI developers to distribute a provenance detection tool with their models, solicit public feedback on that provenance detection system, and conduct extensive adversarial testing on their models for the robustness of their provenance detection system, which would have had to have been shared with the California government. All these requirements would have been particularly burdensome for open-source projects, some of which are maintained by individual graduate students pursuing research.

[30] California Assembly Bill 3211, "California Digital Content Provenance Standards," https://legiscan.com/CA/text/AB3211/id/2984195.

[31] Anna Tong, "OpenAI Supports California AI Bill Requiring 'Watermarking' of Synthetic Content." *Reuters*, August 26, 2024, https://www.reuters.com/technology/artificial-intelligence/openai-supports-california-ai-bill-requiring-watermarking-synthetic-content-2024-08-26/.

Finally, the bill created new mandates for social media companies and other large web properties to identify whether every piece of content shared on their platforms is synthetic, partially synthetic, or human generated. Such an undertaking may not be technically feasible, for reasons explained later in this report.

In summary, although AB 3211 aimed to tackle important issues related to AI-generated content, its ambitious and rigid requirements presented practical difficulties. The bill's focus on watermarking and labeling would likely not have sufficiently addressed the underlying challenges of AI deception and could have inadvertently created new problems, such as stifling technological progress and placing undue burdens on developers.

How Watermarking and Authenticity Technical Standards Work

Ultimately, the provenance and authenticity of digital content is a technological problem, which means that its solution will be—at least in part—technological. The current state of technological solutions, however, reveals serious problems, and whether they can be resolved is unclear. This section will outline the approaches that have been put forth thus far and will explain where they fall short.

Technological solutions to AI-generated content validation can be categorized into either *watermarks* or *metadata*.

Watermarks

Watermarks are modifications to an AI model's outputs that allow those outputs to be recognized as AI generated. Unlike physical watermarks, the watermark on a piece of AI-generated content is not usually noticeable to a human. Instead, it is visible only with a specialized detection algorithm, allowing AI-generated content to be used in professional settings such as a corporate website without an obvious visual detriment. However, in any situation in which it is important to know whether content is human generated or AI generated, these detection algorithms can be employed.[32]

To understand how the watermarks work, one has to understand a bit about how generative AI models work. This section will use language models as an example, but the reader should note that the broad principles apply—with

[32] See, for example, John Kirchenbauer et al., "A Watermark for Large Language Models," May 2024.

differences in the details —to other kinds of generative models, such as image, video, or audio. Such models are not deterministic—that is, they do not output a reliably predictable response, even to the same input. One can ask ChatGPT the same question 100 times and get 100 different replies, even if they are only slightly different. For every word that ChatGPT generates, the model makes a prediction about what word "should" come next based on its training data. In fact, every time ChatGPT generates a word, it is determining the probability for *every single word* in its vocabulary and selecting one. This selection process is, in part, random.[33]

Watermarks work by biasing this selection process in favor of some words over other words. The resultant output from the model will have the same meaning, but its word selection will be subtly altered, and that alteration is the watermark.

The detection algorithm that accompanies a watermark is designed to detect these subtle output alterations. In principle, it can be designed with a high degree of specificity: a watermark can communicate more than simply the fact that the content in question is AI generated. It can indicate what model generated the content or even, in principle, what kind of prompt the model was responding to. Say, for example, a model "suspects" that it has received a prompt to generate an essay for a college student; in principle, a specific watermark for that category of model output could be applied.[34]

Watermarks of this kind are easy for developers to add to existing models because they do not modify the weights—the internals—of the model. Instead, the watermark is implemented through a simple bit of code that tells the model to bias its responses in the specific ways described here. Thus it is easy to "retrofit" watermarks onto existing models, or even to swap out watermarks as more advanced methodologies become available.

This approach has several inherent tradeoffs. First, just as it is easy to add the code that embeds the watermark, that code is similarly easy to remove. This statement is particularly true for open-source or open-weight AI models, whose weights and (sometimes) associated code are made available for download by anyone. Thus, any policy regime in which open-source AI is preserved is one in which it is possible—at least in principle—to remove watermarking code from models. This tradeoff would have to be considered in the broader context

[33] For an in-depth treatment of language model basics, see Alec Radford et al., "Language Models Are Unsupervised Multitask Learners," 2018; Timothy B. Lee and Sean Trott, "Large Language Models, Explained with a Minimum of Math and Jargon," *Understanding AI*, July 27, 2023.
[34] Interview with Professor Scott Aaronson, University of Texas at Austin. July 30, 2024.

of the debate over open and closed models, which is beyond the scope of this document.[35]

Second, watermarks currently can be reliably embedded only over text sequences of certain lengths. Detecting the watermark in an essay-length model output is possible, for example, but not usually in a tweet-length output. This concept is particularly important to understand for laws focused on watermarking content shared on social media; many social media posts are simply not long enough for current watermarks to be detectable.[36]

Third, watermark detection algorithms do not give definitive answers about whether a specific piece of content was AI generated. Instead, they provide probabilities that content came from an AI model. The shorter the content is, the less likely a detection algorithm is to find a high probability that it was AI generated. Thus, in situations in which one needs to make a binary judgment about whether content was or was not AI generated, false negatives and false positives are distinct possibilities.[37]

Fourth, and most important, these watermarks can be destroyed through editing the content in question. If users generate an essay with ChatGPT and change the wording, they may inadvertently destroy the watermark, even if the text retains large amounts of AI-generated words. And if the watermark detection algorithm is made available to the public—which is almost inevitable in any regulatory regime that mandates watermarks—adversarial actors could easily modify their outputs just enough to avoid detection.[38]

Metadata

Metadata is, put simply, data about other data. The date a photo was taken, the author of a file on a computer, and the location from which a social media post was sent are all forms of metadata. Some technological solutions to validating content authenticity involve applying unique metadata to that content. Those

[35] For various assessments on the risks and benefits of open-source AI, see Dean W. Ball, "Free as in Speech or Free as in Beer? Why Open-Source AI Is Essential," *Hyperdimensional,* January 17, 2024; Mark Zuckerberg, "Open Source AI Is The Path Forward," *Meta,* July 23, 2024; Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models," Centre for the Governance of AI, September 29, 2023; National Telecommunications and Information Administration, "Dual-Use Foundation Models with Widely Available Model Weights Report," July 30, 2024.

[36] Aaronson interview.

[37] Siddarth Srinivasan, *Detecting AI Fingerprints: A Guide to Watermarking and Beyond,* Brookings Institution, January 4, 2024.

[38] Sasha Luccioni et al., "AI Watermarking 101: Tools and Techniques," *Hugging Face,* February 26, 2024.

metadata may include both *provenance* and *authenticity* information, answering questions such as “Where did this come from?,” “Who created it?,” “Has this content been modified?,” or “Was this content human or AI generated?” An important consideration is that this approach does not work well for text because in many practical settings, text does not have metadata attached to it. A Microsoft Word document has metadata about that file, but if one simply copies the text in the document into another application, those metadata do not transfer; therefore, reliably attaching metadata to text is infeasible. Thus, metadata-based solutions should best be thought of as applying to images, videos, and audio.

The best-known example of a metadata-based content authenticity standard is the Coalition for Content Provenance and Authenticity (C2PA). C2PA is a technical standard for attaching certain kinds of metadata to digital content and for displaying that content to consumers. Its members include generative AI companies such as OpenAI and Eleven Labs (maker of the most popular AI voice software), large technology platform companies such as Google and Microsoft, social media companies such as TikTok, news outlets such as the BBC, and camera makers such as Nikon and Sony.[39] C2PA can be used to apply metadata to both AI-generated and human-generated content. Examples of C2PA metadata include the following:[40]

- Timestamp: the time the content was created
- Creator information: details about the original creator of the content or about the AI model that created it
- Geolocation data: where the content was created
- Device information: what kind of device captured the content (for example, the make and model of the camera used to take a photograph)
- Edit history: a log of all edits made to the content

The most important pieces of C2PA metadata, however, are the content’s *digital signature* and its *hash value,* unique cryptographic identifiers that verify the content’s integrity and whether it has been modified from the original.

C2PA is envisioned as a soup-to-nuts solution for validating the authenticity of digital information. It can be embedded directly into cameras so that the unique cryptographic information is attached to photos at the time the images

[39] Coalition for Content Provenance and Authenticity, “Membership,” https://c2pa.org/membership/.

[40] Coalition for Content Provenance and Authenticity, “C2PA Specifications,” https://c2pa.org/specifications/specifications/2.0/index.html.

are captured. Generative AI companies can attach C2PA metadata to the photos, videos, and audio their models create, clearly indicating that the content was created by an AI model. The information can be integrated into photo editing software so that all edits made are stored as C2PA metadata. C2PA can be surfaced on news websites or social media platforms so that consumers can easily view authenticity and provenance information. A crucial fact is that content creators retain the ability to keep any part of the C2PA metadata private—for example, if they do not wish to share the location.

When all companies throughout the process of producing content adopt one standard, and when all relevant parties act in good faith, C2PA will play a useful role in helping journalists, courts, and everyday users sort fact from fiction online.

Unfortunately, this situation will not always be the case. First, adoption across the many industries implicated by C2PA will be varied and complex. This problem could be addressed by legal mandates, imposed at the state or federal level, to adopt C2PA. However, beyond the high compliance costs this would entail, such a regulation could also lead to path dependency. Technical standards change over time and can even be replaced; it is a normal and healthy part of technological progress, and laws that freeze standards in place can inadvertently arrest the development and diffusion of new technologies into the market.

Also, however, whether such a law would achieve the desired effect is far from clear because whether C2PA is sufficiently robust to work is not clear. C2PA has been shown by security experts to be exceedingly easy to break.[41] Stripping C2PA metadata from content that is stored on one's own computer is trivial. Indeed, it is easy to do by accident because many current photo editing applications and social media websites do not support C2PA and strip the metadata from the file automatically.

Even if those products do adopt C2PA, however, removing C2PA metadata will remain trivial for a user. Even simplistic methods are fiendishly difficult to counteract: a user can take a screenshot of a picture, for example, and have a duplicate of the original with none of the metadata.

Even more troubling, producing counterfeit C2PA metadata is possible. Although it requires some technical skill, a motivated user can generate a fake image and attach "authentic" C2PA metadata to it. In a policy regime in which social media platforms are required to display C2PA metadata, this feature could

[41] Neal Krawetz, "C2PA's Butterfly Effect," *The Hacker Factor Blog,* November 16, 2023. See also Neal Krawetz, "C2PA from The Attacker's Perspective," *The Hacker Factor Blog,* May 9, 2024.

be used to produce an inauthentic image—say, a deepfake—and have it portrayed on social media as genuine, even human created. Thus, overreliance on C2PA could lead to the worst of both worlds: scenarios in which consumers are led to believe that an image is real when it is, in fact, artificial. Even one such real-world example of this, if it involves a sufficiently high-profile event (for example, the attempted assassination of former President Trump), could quickly erode any public confidence in C2PA, obviating the point of hypothetical regulations to require C2PA and the standard as a whole.

This situation may seem like a design flaw of C2PA, but, in fact, it stems from characteristics of C2PA that are better thought of as features rather than flaws. One of C2PA's key audiences is professional content creators such as photographers, videographers, and the like—in other words, the people who care most about ensuring that their work can be validated as authentic online. That audience rightfully cares about owning the content they create, which means being able to modify the content arbitrarily. It means being able to share exactly what they intend to share and nothing more—either for artistic reasons or for privacy reasons, as outlined above.

At a fundamental level, this capability means that, to garner widespread adoption, C2PA must afford content creators absolute control over their work, including the ability to share it with any audience they choose. Unfortunately, such control is in tension with robustly preserving authenticity and provenance metadata in all circumstances.

Ultimately, digital content, even content with C2PA metadata, is stored as files on computers. So long as that remains true, fundamental limits will exist on how much control third parties can exercise over what can and cannot be done to those files. The only exception to this rule is files with digital rights management (DRM) applied. DRM is commonly used for proprietary content. For example, if a user legally purchases an e-book, that book will be stored as a file, like any other content. DRM is used to ensure that this file can only be read by the user who purchased the book. Of course, such a tool would be inappropriate in the context of C2PA, which is explicitly intended to include professional content creators—in other words, people who have every right to share the content in question.

Given the problems associated with both metadata and watermarks, it is likely too early to mandate their use in a broad range of settings. However, these technical solutions can be used in targeted ways in concert with other policy mechanisms to attain greater confidence about the authenticity of content in contexts in which it matters the most.

A Policy Framework for State-Based Deepfake and Content Authenticity Legislation

Because of the challenges inherent in reliably identifying AI-generated content and the difficulty (perhaps even total infeasibility) of eliminating AI systems that can evade watermarks, policymakers should take a measured stance. The following handful of broad principles flow naturally from the analysis.

Focus on demonstrated harms

Post hoc legislation is far easier to enforce than ex ante legislation and poses a far lower risk of creating mandates that either freeze technology in place or otherwise impede innovation. Fortunately, in most if not all states, the legal framework is already in place: fraud and defamation are already crimes and, at most, such statutes will need to be updated to resolve any ambiguities related to their applicability to generative AI.

Post hoc legislation need not be devoid of any mandates on companies. For example, requiring social media platforms to have a process for removing malicious synthetic content at the request of a victim and a timeline for making a decision on such removals is a reasonable measure. These requirements should be minimalistic, straightforward, and imposed only on the largest social media platforms to avoid creating a patchwork of regulations or an undue compliance burden on small firms. Harmonizing such requirements, where possible, with similar laws in other states is also wise.

Specific recommendations include the following:

- Convene a statewide commission or working group, chaired by the attorney general, to review existing state laws for ambiguities, gaps, loopholes, and other flaws relating to generative AI, with a deadline for that group to issue recommendations within one year.
- Update existing law as appropriate.
- Do not impose fines or liability on social media companies or other technology platforms, such as phone service providers, through which deceptive AI-generated content can be transmitted or created; focus enforcement on malicious actors.
- Create a safe harbor—protections from legal liability—for social media platforms, websites, and other hosts of potentially malicious AI-generated content, provided that they comply with takedown requests.

- Mandate a reasonable timeline (say, 10 business days) for platforms and websites to remove content, predicating the safe harbor provision on compliance with this timeline.
- Ensure that standard First Amendment exceptions are clearly made (newsworthy events, satire, political speech, etc.).
- Predicate laws on proving malicious intent on the part of the person who distributed the deepfake image; do not make it a crime to make a deepfake, which could end up being overly broad. Instead, tailor the statute narrowly to the distribution of a digital replica of another person with malicious or deceptive intent.
- Extend protections on identity to a reasonable time frame after an individual's death—no more than five years.

The framework offered here will grant all state residents robust protections against malicious digital replicas without impeding innovation in AI or overburdening other industries.

Mandate technical solutions in targeted applications

Metadata and watermarks are not ready for broad-based applications—for example, mandating watermarks on all content on social media would be premature. Laws that seek to achieve this outcome are likely to stifle innovation, cause unintended consequences, and even create a false sense of security about the authenticity of content. Instead, states should concentrate on using these tools where they matter most: documents submitted by individuals and businesses to the state under penalty of perjury.

Specific recommendations include the following:

- Mandate that state agencies and courts use watermark detection systems, as they become available, to check the authenticity of content submitted to the state under penalty of perjury.
- Understand that these detection systems may lead to false negatives and false positives; do not treat their judgments as dispositive but, instead, as a prompt for further investigation and inquiry.
- Allow state agencies and courts substantial flexibility in determining what specific detection system to use. This area is likely to advance rapidly over the coming years, so statutes should not mandate any specific system.

- Put this mandate into effect with a substantial delay (beginning in 2026 or 2027) to give the industry time to create better detection systems. Although such systems are on the horizon, current detection systems are quite poor.
- Require anyone submitting a document under penalty of perjury to confirm whether any of the content is synthetically generated; the use of synthetic content should not per se be considered a violation of the law unless (a) it is materially demonstrated to be a known falsification or (b) the submitter is demonstrated to have lied about whether the content was synthetically generated.
- Create a statewide commission to study the effect of generative AI on evidentiary standards in state courts; have that commission consider means of educating judges, prosecutors, and other relevant state and local government employees about how to identify synthetic content, particularly images and video.

These solutions will not be a silver bullet; likely, no silver bullet exists for this problem. Nonetheless, taken together they will give a means of redress to state residents who are the victim of AI-related fraud or defamation, and they will give the state government the information it needs to continue performing its vital functions.

Conclusion

As with many areas of AI, there are currently more open questions than answers. Society does not know if AI's current rather muted effect on the information environment is merely the calm before a coming storm. We do not know how far technological solutions to this inherently sociotechnical problem will take us. What society does know, however, is that AI progress is likely to be both rapid and uncertain. It is not a problem to be solved; it is a fact with which to reckon.

What this fact means is that policymakers should seek to gain information and insight while retaining optionality and flexibility. This stance, in turn, means that legislative mandates of safety or epistemic certainty—tempting though they may be—are likely to create more problems than they solve. Instead, knowledge, familiarity with AI tools, and an efficient means for citizens who have been harmed to seek relief will serve states far better. As ever, then, state policymakers should keep a watchful eye on developments in AI yet avoid the urge to rush to legislative judgment.

About the Author

Dean Woodley Ball is a research fellow in the Artificial Intelligence & Progress Project at George Mason University's Mercatus Center and creator and author of *Hyperdimensional*, a Substack newsletter. His work focuses on emerging technologies and the future of governance. He has written on topics including artificial intelligence, the future of manufacturing, neural technology, bioengineering, technology policy, political theory, public finance, urban infrastructure, and prisoner reentry.